State-of-the-art generalisation research in NLP:
a taxonomy and review

Dieuwke Hupkes[∞], Mario Giulianelli[◇], Verna Dankers[∞+], Mikel Artetxe[∞]
Yanai Elazar[⊥⋆], Tiago Pimentel[◁], Christos Christodoulopoulos[⊗], Karim Lasri[∧]
Naomi Saphra[†], Arabella Sinclair[%], Dennis Ulmer[∪], Florian Schottmann[⊙]
Khuyagbaatar Batsuren[□], Kaiser Sun[∞], Koustuv Sinha[∞], Leila Khalatbari[~]
Rita Frieske[~], Ryan Cotterell[⊙], Zhijing Jin[∞]

dieuwkehupkes@fb.com m.giulianelli@uva.nl
vernadankers@gmail.com

[∞]Meta AI [◇]University of Amsterdam [+]University of Edinburgh [⊗]Amazon Alexa AI
[◁]University of Cambridge [†]NYU [∧]École Normale Supérieure-PSL [⊥]Allen Institute of AI
[⋆]University of Washington [%]University of Aberdeen [∪]IT University of Copenhagen [⊙]ETH Zurich
[□]National University of Mongolia [~]Hong Kong University of Science and Technology

Abstract

The ability to generalise well is one of the primary desiderata of natural language processing (NLP). Yet, what 'good generalisation' entails and how it should be evaluated is not well understood, nor are there any common standards to evaluate it. In this paper, we aim to lay the groundwork to improve both of these issues. We present a taxonomy for characterising and understanding generalisation research in NLP, we use that taxonomy to present a comprehensive map of published generalisation studies, and we make recommendations for which areas might deserve attention in the future. Our taxonomy is based on an extensive literature review of generalisation research, and contains five axes along which studies can differ: their main motivation, the type of generalisation they aim to solve, the type of data shift they consider, the source by which this data shift is obtained, and the locus of the shift within the modelling pipeline. We use our taxonomy to classify over 400 previous papers that test generalisation, for a total of more than 600 individual experiments. Considering the results of this review, we present an in-depth analysis of the current state of generalisation research in NLP, and make recommendations for the future. Along with this paper, we release a webpage where the results of our review can be dynamically explored, and which we intend to update as new NLP generalisation studies are published. With this work, we aim to make steps towards making state-of-the-art generalisation testing the new status quo in NLP.

1 Introduction

Good generalisation, roughly defined as the ability to successfully transfer representations, knowledge, and strategies from past experience to new experiences, is one of the primary desiderata for models of natural language processing (NLP), as well as for models in the wider field of machine learning (Elangovan et al., 2021; Kirk et al., 2021; Lake et al., 2017; Linzen, 2020; Marcus, 2018, 1998; Schmidhuber, 1990; Shen et al., 2021; Wong and Wang, 2007; Yogatama et al., 2019, i.a.). For some, generalisation is crucial to ensure that models behave robustly, reliably, and fairly when making predictions about data different from the data that they were trained on, which is especially valuable when models are employed in the real world. Others see generalisation as directly equivalent to good performance and believe that without it a model does not truly conduct the task we intended it to. Yet others strive for good generalisation in models because they believe models should behave in a human-like way – and humans are known to generalise well. While the importance of generalisation is almost undisputed, and there are countless papers on the matter, systematic generalisation testing is not the status quo in the field of NLP. At the root of this problem lies the fact that there is little understanding and agreement about what good generalisation actually entails, and what types of generalisation should be prioritised in which scenarios. While generalisation is widely discussed in NLP – in the past five years, in the ACL anthology alone over 1200 papers mentioned it in their title or abstract – there exists no systematic framework to characterise and discuss generalisation. Different studies differ amply in the assumptions they make about when and how models should generalise, and they use a wide range of different experimental and evaluation setups. As a result, it is difficult to understand what the current state of the field is when it comes to generalisation. It is difficult to understand how results in this area relate to each other, what sorts of generalisation are being addressed and which are neglected, which forms of generalisation testing we should prioritise, and how we can adequately assess generalisation in the first place. Missing answers to all of those questions are standing in the way of better model development: what we cannot measure, we cannot improve.

In this paper, we introduce a new framework to systematise and understand generalisation research, and we address questions like the ones above. More precisely,

i) We *design a taxonomy to characterise generalisation research*, grounded in hundreds of existing generalisation studies;

ii) We *present an in-depth analysis* based on over 400 papers with generalisation experiments that have come out in the last decades;

iii) *We make recommendations* for which areas we believe deserve attention in the near future and;

iv) We *release a set of online tools* that can help readers to better understand the current landscape of generalisation-testing, exploring the data by themselves.

With our taxonomy, analysis and online tools, we aim to **lay the groundwork for making *state-of-the-art generalisation testing* the status quo in NLP**.

1.1 What is generalisation?

Broadly speaking, generalisation is evaluated by assessing how well a model performs on a test dataset, given the relationship of this dataset with the data the model was trained on. For decades, it was common to put only one simple constraint on this relationship: that the train and test data are different. Typically, this was achieved by randomly splitting available data into a training and a test partition. Generalisation was, thus, evaluated by training and testing models on different but similarly sampled data, assumed to be independent and identically distributed (i.i.d.). In the past 20 years, we have seen great strides on such random train–test splits in a range of different applications. Since the first release of the Penn Treebank (Marcus et al., 1993), F_1 scores for labelled constituency parsing went from values in the high 80's at the end of the previous century (Collins, 1996; Magerman, 1995) and the first ten years of the current one (e.g. Petrov and Klein, 2007; Sangati and Zuidema, 2011) to scores up to 96 in the recent past (Mrini et al., 2020; Yang and Deng, 2020). On the same corpus, performance for language modelling went from per-word perplexity scores well above 100 (Kneser and Ney, 1995; Rosenfeld, 1996) to a score of 20.5 in 2020 (Brown et al., 2020). In many areas of NLP, the rate of progress has become even faster in the last few years. Scores for the popular evaluation set GLUE went from values between 60 and 70 at its release (Wang et al., 2018), to scores exceeding 90 less than a year after (most famously, Devlin et al., 2019), with performances on a wide range of tasks reaching and surpassing human-level scores (e.g. Devlin et al., 2019; Liu et al., 2019b; Wang et al., 2019, 2018). Yet more recently, strongly scaled-up models (e.g. Chowdhery et al., 2022) showed astounding performances on almost all existing i.i.d. natural language understanding benchmarks.

With this progress, however, came the realisation that, for an NLP model, reaching very high or human-level scores on an i.i.d. test set does not imply that the model robustly generalises to a wide range of different scenarios. In the recent past, we witnessed a surge of different studies pointing out generalisation failures in neural models that have state-of-the-art scores on random train–test splits (Blodgett et al., 2016; Khishigsuren et al., 2022; Kim and Linzen, 2020; Lake and Baroni, 2018; Marcus, 2018; McCoy et al., 2019; Plank, 2016; Razeghi et al., 2022; Sinha et al., 2021, to give just a few examples). Some show that when models perform well on i.i.d. test splits, they might rely on simple heuristics that do not robustly generalise in a wide range of non-i.i.d. scenarios (Gardner et al., 2020; Kaushik et al., 2019; McCoy et al., 2019), that models over-rely on stereotypes (Parrish et al., 2022; Srivastava et al., 2022), or bank on memorisation rather than generalisation (Lewis et al., 2021; Razeghi et al., 2022). Others, instead, discuss cases in which performances drop when the evaluation data differs from the training data in terms of genre, domain or topic (e.g. Malinin et al., 2021; Michel and Neubig, 2018; Plank, 2016), or when it is produced by different subpopulations (e.g. Blodgett et al., 2016; Dixon et al., 2018). Yet others focus on models' inability to generalise compositionally (Dankers et al., 2022; Kim and Linzen, 2020; Lake and Baroni, 2018; Li et al., 2021b), structurally (Sinha et al., 2021; Weber et al., 2021; Wei et al., 2021), to longer sequences (Dubois et al., 2020; Raunak et al., 2019), or to slightly different task formulations of the same problem (Srivastava et al., 2022).

The examples above are just a few in a long list of studies that aim to investigate the generalisation abilities of NLP models, focusing in particular on models and training regimes that score well on traditional train–test splits. At the same time, these works differ amply in the assumptions they make about when and how models should generalise, and the evaluation settings they use to evaluate that. They encompass a wide range of generalisation-related research questions, and they use a wide range of different methodologies and experimental setups. Taken together, this body of work thus illustrates that there is no real agreement on what kind of generalisation is important for NLP models, and it also brings into question what kind of generalisation capabilities recent breakthroughs actually reflect. How should generalisation be tested for, if not with i.i.d. splits? How do we discover which types of generalisation should be prioritised, how the results of different studies relate to each other, what types of generalisation are already well addressed and which are neglected? Ultimately, on a more meta-level, how can we make progress on these important questions without a systematic way to discuss generalisation in NLP?

1.2 The generalisation taxonomy: a bird's eye view

It is exactly this meta-question that we aim to address with this paper, by proposing a framework that can be used to systematically characterise and understand generalisation research. More specifically, we present a **generalisation taxonomy**, an **analysis** of existing work on generalisation, and a **set of online tools** that can be used by researchers to explore and better understand generalisation studies in NLP. The generalisation taxonomy we propose is based on a detailed analysis of a large number of existing studies on generalisation in NLP, and it includes the main five axes along which those studies differ.[1] The five axes capture different aspects of generalisation studies, that together form a comprehensive picture of the motivation and goal of the study and provide information on important choices in the experimental setup.

The first axis of our generalisation taxonomy (§2) is the high-level **motivation** for the study. The motivation of a study impacts or even determines what type of generalisation is desirable, as well as what kind of conclusions can be drawn from a model's display or lack of generalisation. Furthermore, the motivation of a study shapes its experimental design. It is therefore important for researchers to be explicitly aware of it, to ensure that the experimental setup aligns with the questions they seek to answer.

[1] An graphical representation can be found in Figure 1.

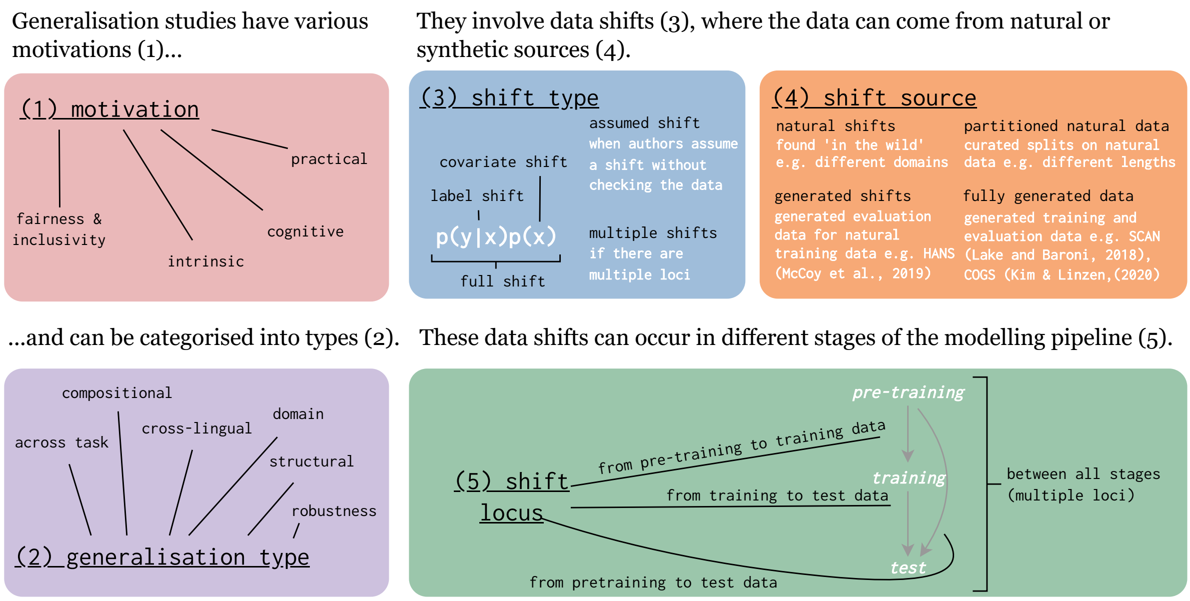


Figure 1: A graphical representation of the NLP generalisation taxonomy we present in this paper. The taxonomy consists of five different (nominal) axes, that describe the high-level *motivation* of the work (§2); the *type* of generalisation the test is addressing (§3); what kind of *data shift* occurs between training and testing (§4), and what the *source* and *locus* of this shift are (§5 and §6, respectively).

We consider four different types of motivations: the *practical* motivation, the *cognitive* motivation, the *intrinsic* motivation, and the *fairness and inclusivity* motivation.

The second axis in our taxonomy (§3) indicates the **type of generalisation** the test is addressing. This axis describes on a high level what exactly it is that a generalisation test is intended to capture, rather than considering why or how, making it one of the most important axes of our taxonomy. In the literature, we have found six main types of generalisation: *compositional* generalisation, *structural* generalisation, *cross-task* generalisation, *cross-lingual* generalisation, *cross-domain* generalisation, and *robustness* generalisation.

The third axis in our taxonomy (§4) describes what kind of **data shift** is considered in the generalisation test. This axis adds a statistical interpretation to our taxonomy and derives its importance from the fact that data shift plays an essential formal role in defining and understanding generalisation from a statistical perspective, as well as from the fact that different types of shifts are best addressed with different kinds of experimental setups. On the data shift axis, we consider three shifts which are well-attested in the literature: *covariate shift*, *label shift* and *full shift*. We further include two additional types of shift – *assumed shift* and *multiple shifts* – to account for studies that cannot be labelled with any of the three main shift types.

In the fourth axis of our taxonomy (§5), we consider what is the **source** of the data shift used in the experiment. The source of the data shift determines how much control the experimenter has over the training and testing data and, consequently, what kind of conclusions can be drawn from an experiment. We distinguish four different sources of shifts: *naturally occurring shifts*, *artificially partitioned natural corpora*, *generated shifts* and *fully generated datasets*.

In the last axis of our taxonomy (§6), we consider what is the **locus** of the data shift, or, in other words, for what part of the modelling pipeline generalisation is investigated. The locus of the shift, together with the shift type, forms the last piece of the puzzle, as it determines what part of the modelling pipeline is investigated and thus the kind of generalisation question that can be asked. On this axis, we consider shifts between all stages in the contemporary modelling pipeline – pretraining, training and

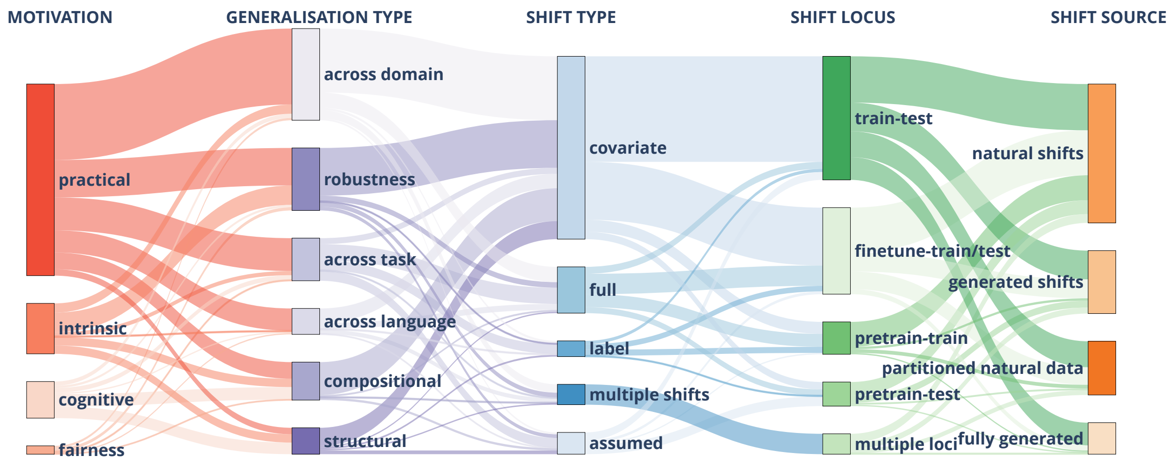


Figure 2: An overview figure of our literature review, including interactions. An interactive version of this plot can be found on our website `https://genbench.github.io/visualisations`. For more detailed explanations and analyses, we refer to §7.

testing – as well as studies that consider shifts between multiple stages simultaneously.

1.3 Our review and analysis: a sneak-preview

Using our taxonomy, we conduct an extensive literature review, in which we survey over 400 papers in the ACL anthology that contain the (sub)words generali(s|z)ation or generali(s|z)e in their title or abstract and that consider some form of data shift in their experiment. Using different visualisations, we analyse the most relevant trends and find several noteworthy patterns (§7.2.2).

First, we observe that *the experimental design of a study is not always lined up with its motivation*. To give an example, several studies considering compositional generalisation from a practical perspective use generated data not reflective of the scenarios that models might in practice be employed in, making it difficult to draw conclusions that match the proposed motivation of the study. As such, this demonstrates the importance of the motivation axis in designing generalisation studies. Then, we find that an increasing number of papers investigating generalisation *does not explicitly consider the relationship between train and test data*. This trend is likely due to the computational and engineering advances that allow model training on extremely large corpora: the ever-growing sizes of the training corpora, which are furthermore often not in the public domain, make it increasingly difficult to determine the relationship between train and test data, and consequently how generalisation should be evaluated in these scenarios. A similar issue arises in the setup where pretrained models are tested without further finetuning, such as in prompting or in-context learning setups. In such setups, there is a shift between pretraining and testing, which is – for the same reasons as laid out above – difficult to analyse. Our taxonomy provides the means to understand these problems, and it illustrates that they require further thought in the future to allow for generalisation testing in such increasingly popular setups. A third important observation is that many *papers that contain a multi-stage modelling pipeline investigate generalisation in one part of that pipeline, but not in the other* (as can be seen in Figure 2, by comparing the number of pretrain-train and finetune train–test loci with the number of multiple loci). For instance, a researcher might extensively evaluate whether a pretrained model can be finetuned on a large number of tasks, but use random splits to assess each individual task, or, conversely, they might test generalisation in the finetuning stage for a single task and draw conclusions about the pretrained model, without considering whether those results hold also when the model is finetuned on other tasks. Both these scenarios lead to models that generalise suboptimally when considered as a whole. Therefore, we argue that in the future it is important to prioritise models that generalise well at all levels of the modelling pipeline, and not just in one phase. Another takeaway is that our results suggest that *more meta-studies might be needed that compare results across different values of the same axis*, for example, to understand what is the relationship between results obtained with fully generated data and generated shifts. Such studies can improve our understanding of how different experimental design choices impact the conclusions that can be drawn from an experiment. Lastly, we find that both *studies on cross-lingual generalisation and studies with a fairness motivation are under-represented* in our review. In part, this may indicate that such studies refer less explicitly to generalisation in their title and abstract. However, we hypothesise that also prioritisation in the field plays a role. In particular, the fact that NLP is very English-centric (e.g. Bender, 2011; Cotterell et al., 2018) is likely to impact the number of cross-lingual studies. For fairness, on the other hand, under-representation could stem from the fact only relatively recently awareness of the potential harmfulness of models trained on large, uncontrolled corpora has started to grow. Either way, we believe that both cross-lingual generalisation and fairness are important matters to prioritise in the future. We also call to the reader to propose existing papers with these axis values via our website, so that we can increase our coverage.

1.4 Outline and contributions

We believe that generalisation testing **should be the new status quo in NLP**, and with this work, we aim to **lay the groundwork** for **making that a reality**. In summary, the contributions of our work are the following:

i) We present an axis-based generalisation taxonomy that can be used to characterise generalisation studies in NLP;

ii) We review 449 papers, containing a total of 619 generalisation experiments, using this taxonomy;

iii) With these survey results, we discuss the status of generalisation research in NLP, and we provide suggestions to steer the field towards more sound and exhaustive generalisation tests.

iv) We present a website where our review results can be (visualised and textually) explored and (new) generalisation studies can be incorporated.

In the remainder of this paper, we will first discuss the different axes of our taxonomy in more detail (§2-6). After that, in §7, we will present our review and analysis of the current state of generalisation research. In §8, we wrap up by summarising our main findings and making concrete recommendations for the future.

2 Motivation: what is the high-level motivation for a generalisation test?

Now that we have outlined our main objectives, we discuss the five axes in our proposed taxonomy. The first axis we consider is the high-level motivation of a generalisation study. We identified four closely intertwined goals of generalisation research in NLP, which we refer to as the *practical*, the *cognitive*, the *intrinsic*, and the *fairness* motivation. The motivation of a study impacts or even determines what type of generalisation is desirable, as well as what kind of conclusions can be drawn from a model's display or lack of generalisation. Consider, for instance, cases in which humans fail to generalise. For a study with a cognitive motivation, model failures in such cases might not be problematic, or perhaps even desirable. This is unlikely to be the case for studies with a fairness or practical motivation, where propagation of human biases is usually problematic. Connected to this, the motivation of a study shapes the decisions that need to be made for its experimental design. It is therefore important for researchers to be explicitly aware of it, to ensure that the experimental setup aligns with the questions they seek to answer. For a study with a practical motivation, for example, it is typically important to consider a data setup that matches real-world scenarios a model might occur in; this is less relevant for studies considering generalisation with a cognitive or intrinsic motivation. Given its strong influence on the other axes of the taxonomy, a study's high-level motivation is the first axis we discuss. We describe the four motivations we distinguish below.[2]

Practical: in what settings can the model be used or improved? One frequently posed motivation to study generalisation is of a highly practical nature. Studies that consider generalisation from a practical perspective seek to assess in what kind of scenarios a model can be used, or focus on improving model generalisation. One question that is often addressed with a primarily practical motivation is how well models generalise to different domains or differently collected data. For instance, Michel and Neubig (2018) consider how well machine translation models trained on canonical text can generalise to noisy data from an internet platform, and Lazaridou et al. (2021) investigate language model generalisation to different time periods. Other questions that are frequently addressed from a practical perspective concern biases in the training data, and whether models robustly generalise to datasets that do not share these (spurious) biases (e.g. Behnke et al., 2022; Zhou et al., 2021).

Cognitive: does the model generalise like a human? A second high-level motivation that drives generalisation research is cognitively oriented and can be separated into two underlying categories. The first category is related to model behaviour: human generalisation is a useful reference point for the evaluation of model generalisation in NLP, because human generalisation is known to be powerful (e.g. Lake et al., 2017; Marcus, 2003) and, perhaps more importantly, precisely the type of generalisation that is required to successfully model natural language. Humans learn quickly, from fewer data than models, and they easily (compositionally) recombine concepts they already know to understand concepts they have never before encountered (Fodor and Pylyshyn, 1988; Linzen, 2020; Marcus, 2018). These feats are arguably also important for models; they, therefore, provide a good point of reference for generalisation testing.[3] In some cases, it might be difficult to distinguish cognitive and practical motivations: assuming human generalisation is strong, a model that generalises like a human should score well also on practically motivated tests. In our axes-based taxonomy, the difference between *cognitive* and *practical* resides mostly in the types of scenarios that are considered in tests: are the scenarios artificially created to get a higher-level, isolated impression of how their behaviour compares to human-like generalisation, or are the scenarios realistic and practically relevant?

The second, more deeply cognitively inspired category contains work that evaluates generalisation in models to learn more about cognition and language (e.g. Baroni, 2021; Hupkes, 2020; Marcus, 1999; McClelland and Plaut, 1999). Studies in this category investigate whether a particular model generalises primarily to derive new hypotheses about how human generalisation might work. For instance, Lakretz et al. (2021b) perform a detailed study of how LSTM models generalise to specific kinds of nested syntactic constructions, which they then use to inform a human experiment on the same syntactic constructions.

[2] As we will see in what follows, the same questions can often be asked with different underlying motivations. This makes it sometimes difficult to identify what exactly the motivation of a generalisation study is. Often, studies may inform conclusions along all four dimensions. However, given the importance of the motivation for the implications and design of the study, we nevertheless try to identify the main guiding motive of a study in our review in §7, and we encourage researchers to be explicit about the motivation of their future studies.

[3] We do not always expect from a model the same type or level of generalisation a human exhibits. There are cases in which it is desirable for models to generalise better than humans, for example across languages – something humans above a certain age typically do not excel at. In other cases, models already generalise better than humans – consider, for instance, a language identification system – and would hardly be useful if they did not.

Intrinsic: does the model capture the task correctly? A third motivation to evaluate generalisation in NLP models, which cuts through the two previous motivations, appertains to the question "*did a model learn the task we intended it to learn, as we intended it to learn it?*". The assumption underpinning this type of research as a whole is that if a model has truly learned the task it is trained to do, it should be able to execute this task also in settings that differ from the exact training scenarios. What changes across studies is the set of conditions under which a model is considered to have appropriately learned a task. For instance, researchers studying compositional generalisation (see §3.1) assume that a correct understanding of language implies that the assumed compositional structure of language is captured. Under that assumption, a model should not have trouble generalising to new inputs that are generated using the same compositional system. Others instead assume that true language understanding implies being able to use language across a wide variety of tasks (see §3.3). Yet others argue that if a model truly captures the relationship between two sentences in NLI tasks (e.g. Bowman et al., 2015a; Marelli et al., 2014; Williams et al., 2018), it should be able to do so across different data sets, even if those were sampled in a slightly different way (e.g. Talman and Chatzikyriakidis, 2019). In studies that consider generalisation from this perspective, generalisation failures are taken as proof that the model – in fact – did not learn the task as we intended it to learn it (but instead showed behaviour that made us think it did, for instance by relying on spurious patterns or non-generalisable heuristics). Furthermore, studies with an intrinsic motivation are usually guided by the purely scientific motive of increasing knowledge and understanding, rather than targeting a specific goal.

Fairness and inclusivity: does the model generalise in a fair and responsible way? A last yet very important motivation for generalisation research is the desire to have models that are fair, responsible and unbiased. One category of studies driven by these concepts, often ethical in nature, asks questions about how well models generalise to diverse demographics, typically considering minority or marginalised groups (e.g. Bender et al., 2021; Blodgett et al., 2016; Koh et al., 2021), or investigates to what extent models perpetuate (undesirable) biases learned from their training data (e.g. Dixon et al., 2018; Hutchinson et al., 2020; Sheng et al., 2019). Another line of research related to both fairness and inclusivity focuses on efficiency, both in terms of the amount of data that is required for a model to converge to a solution as well as the necessary amount of compute. In such studies, efficiency is seen *as a correlate* of generalisation: models that generalise well should learn more quickly and require fewer data (see, e.g. Marcus, 2018). The relationship of efficiency with fairness, inclusivity and responsibility stems from the idea that models that generalise well from small amounts of data are more inclusively applicable – for instance for low-resource languages or demographic groups for which little data is available. Furthermore, models that require less compute are more accessible for groups with smaller computational resources and have a lower environmental impact (see, e.g. Strubell et al., 2019). While we have not mentioned them before in the respective categories, studies on learning efficiency can, naturally, also be motivated by practical concerns, as well as by cognitive interests (e.g. comparing human's and model's sample efficiency).

3 Generalisation type: what type of generalisation is a test addressing?

A second important dimension when it comes to characterising generalisation research is what type of generalisation a test aims to evaluate. The second axis in our taxonomy describes, on a high level, what it is that a generalisation test intends to capture – rather than considering why or how – making it one of the most important axes of our taxonomy. We identify and describe six types of generalisation that are frequently considered in the literature. Some types are rooted in knowledge about human generalisation, such as those that target *compositional* (§3.1) or *structural* generalisation (§3.2). Others, instead, are motivated by more practical concerns, such as generalisation *across tasks* (§3.3), *languages* (§3.4) and

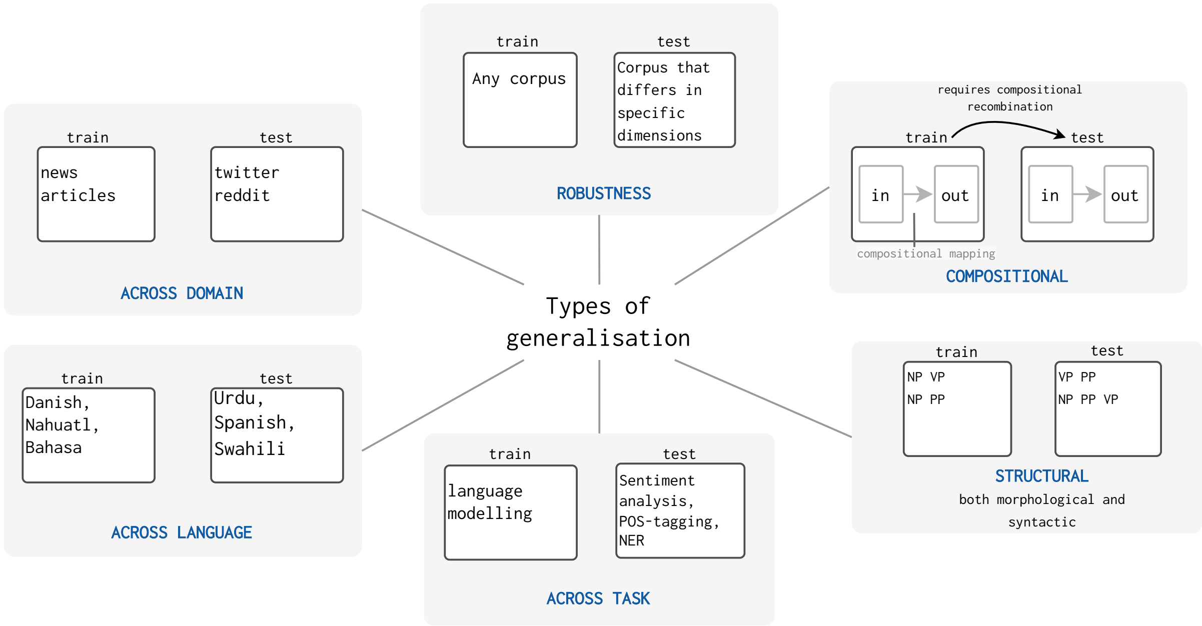


Figure 3: An infographic that illustrates the six different types of generalisation that we consider in our taxonomy, which are explained in more detail in §3.1-§3.6.

domains (§3.5), or by an interest in analysing how *robustly* models generalise (§3.6). An overview of the types we consider is presented in Figure 3.

3.1 Compositional generalisation

The first prominent type of generalisation that can be found in the literature is *compositional generalisation*, which is often argued to underpin human's ability to quickly generalise to new data, tasks and domains (Fodor and Pylyshyn, 1988; Lake et al., 2017; Marcus, 2018; Schmidhuber, 1990). Because of this strong connection with humans and human language, work on compositional generalisation often has a primarily cognitive motivation, although practical concerns such as sample efficiency, quick adaptation and good generalisation in low-resource scenarios are frequently mentioned as additional or alternative motivations (Chaabouni et al., 2021; Linzen, 2020, to give just a few examples). While it has a strong intuitive appeal and clear mathematical definition (Montague, 1970), compositional generalisation is not easy to pin down empirically. Here, we follow Schmidhuber (1990) in defining compositionality as the ability to systematically recombine previously learned elements to map new inputs made up from these elements to their correct output.[4] In language, the inputs are 'forms' (e.g. phrases, sentences, larger pieces of discourse), and the output that they need to be mapped to is their meaning or interpretation. Because of the need for both an input and output space, compositional generalisation is usually evaluated in tasks such as sequence classification (e.g. Bowman et al., 2015b; Hupkes et al., 2018; Veldhoen et al., 2016), machine translation (e.g. Dankers et al., 2022; Liu et al., 2021; Raunak et al., 2019), semantic parsing (e.g. Finegan-Dollak et al., 2018; Keysers et al., 2019; Kim and Linzen, 2020; Shaw et al., 2021) or other kinds of generation tasks (e.g. Hupkes et al., 2020; Lake and Baroni, 2018). In such tasks, the in- and output spaces are clearly distinct, and outputs can straightforwardly be viewed as an interpretation or (proxy) of meaning of its corresponding input. As far as we know, there have not yet been many explicit systematic attempts to evaluate compositionality in (ungrounded)

[4]For an elaborate account of the different arguments that come into play when defining and evaluating compositionality for a neural network, we refer to Hupkes et al. (2020).

language models.[5] If and how compositionality can be adequately evaluated in such models, where the input and output (form and meaning) are conflated in one space (the space defined by the language vocabulary), is a question that is yet to be answered.[6]

3.2 Structural generalisation

Another category of usually cognitively inspired generalisation instead focuses on the extent to which models can produce or generate structurally (grammatically) correct forms, rather than on whether they can assign them correct interpretations. Unlike compositional generalisation, structural generalisation does not require an output space (the meaning or interpretation space; see §3.1). This makes it more straightforwardly evaluated in form-only models (i.e. language models) and completely natural setups (i.e. with no need for simplified synthetic input and output spaces). We distinguish two broad categories of structural generalisation: syntactic generalisation, and morphological generalisation.

Syntactic generalisation Some structural generalisation studies focus specifically on *syntactic generalisation*. They consider whether models can generalise to novel syntactic structures or novel elements in known syntactic structures. For instance, Jumelet et al. (2021) and Weber et al. (2021) filter out from the training data specific licensing environments for negative polarity items and then test whether models nevertheless learn to generalise to such environments. It is unfortunately difficult to conduct this type of study, which involves several different training corpora, using very large language models. On the one hand, their high training cost makes the necessary experiments computationally extremely expensive. On the other hand, generating specific test splits given knowledge of what is in the training data is often also not possible for such models, because their training data is not in the open domain. These limitations prevent researchers from controlling the relationship between the evaluation and training data, and they make it hard to assess to what extent the incidental examples reported for the large language models (most notably, in their respective release papers) are reflective of successful generalisation and, if so, what that entails. Interesting exceptions are a few studies that do explicitly consider shifts between training and testing in the context of syntactic generalisation, such as those presented by Wei et al. (2021), Razeghi et al. (2022), and Elazar et al. (2022). Wei et al. (2021), in particular, investigate how the performance of pretrained language models in tests that assess syntactic rule learning is affected by a term's training data frequency, by varying those frequencies in the training corpus. Razeghi et al. (2022), instead, focus on a larger model trained on more data, and while they do not systematically vary the training corpus, they do an elaborate analysis of how test performance in their trained models (GPT-J and GPT-Neo) is affected by absolute and relative frequencies of specific terms in the model's training data. Even more recently, Elazar et al. (2022) studies the causal effect of simple statistics from the training data, such as co-occurrences, on models' prediction.

Note that the vast majority of other studies focusing on the syntactic abilities of language models (e.g. Giulianelli et al., 2018; Jumelet and Hupkes, 2018; Linzen et al., 2016; Warstadt et al., 2019, 2020) focus on whether and how models recognise, represent, and process syntactic information, or they try to discern the causal mechanisms by which models use such abilities (Amini et al., 2022; Elazar et al., 2021a; Feder et al., 2021). These works do not (explicitly) consider the relationship between the data they test on and the data that a model was trained on, and as such they do not specifically study the models' generalisation abilities across syntactic structures. We will not further discuss these studies, but

[5] There are, however, several studies that focus on *structural* generalisation in such models. Contrary to compositional generalisation, structural generalisation does not focus on the ability of models to correctly interpret new inputs, or assign meanings to them, but only on whether they can generalise to their correct form. We will discuss structural generalisation in the next subsection.

[6] An interesting example to consider in this context is the qualitative study conducted by Brown et al. (2020) to test if GPT-3 can use novel words correctly in a sentence; as another example, a bit further away from traditional forms of compositionality, Talmor et al. (2020) finetune pretrained masked language models on multi-hop composition in question answering.

in our map of generalisation literature (§7), we will include a few papers in which there is an implicit yet clear assumption that the test data substantially differs from the training data, for instance because it includes sentences created with semantically nonsensical words (Gulordava et al., 2018), or unusually deep levels of recursion (Lakretz et al., 2021a,b) that are not likely to naturally occur in corpora.

Morphological generalisation A second category of structural generalisation studies focuses on morphological inflexion, a popular testing ground for questions about human generalisation. Papers focusing on morphological inflexion (e.g. Corkery et al., 2019; Dankers et al., 2021; Kirov and Cotterell, 2018; Liu and Hulden, 2022; Malouf, 2017; McCurdy et al., 2020) are typically rooted in strong cognitive motivations. While most of this work considers i.i.d. train–test splits, recent studies have focused on how morphological transducer models generalise across languages (e.g. McCarthy et al., 2019; Pimentel et al., 2021a; Vylomova et al., 2020) as well within each language (Calderone et al., 2021; Li and Wilson, 2021; Liu and Hulden, 2022; Pimentel et al., 2021b; Szolnok et al., 2021; Wilson and Li, 2021). In doing so, they often take inspiration from *wug* tests, which are used in psycholinguistics to probe morphological generalisation to novel words in humans (Berko, 1958; Marcus et al., 1995). In principle, such studies could also be conducted with large language models but the lack of access to their training data is, again, a complication for determining whether the supposedly novel words were truly never seen by the models.

3.3 Generalisation across tasks

A third and completely different direction of generalisation research considers the ability of a single model to adapt to multiple NLP problems. We refer to this ability as generalisation across tasks, or cross-task generalisation. Along with the great advancements in NLP models, in the past ten years, the nature of cross-task generalisation tests has quite substantially changed; we discuss this evolution in the present section.

Multitask learning Cross-task generalisation in NLP has been traditionally strongly connected to transfer and multitask learning (Collobert and Weston, 2008). In multitask learning, a model is either trained on a set of tasks and evaluated on those same tasks, or pretrained on some tasks and then adapted to others. As this setup favours approaches that benefit from positive transfer across tasks, it implicitly studies forms of cross-task generalisation.[7] Examples of benchmarks that were originally meant to address this kind of cross-task transfer – although they are not used as such any longer – are multitask benchmarks such DecaNLP (McCann et al., 2018), GLUE (Wang et al., 2018) and its successor SuperGLUE (Wang et al., 2019). In recent times, a common approach has been to formulate all tasks as sequence-to-sequence problems, a direction explored in the DecaNLP benchmark (McCann et al., 2018), as well as in modelling, by T5 (Raffel et al., 2020), exT5 (Aribandi et al., 2022) and UnifiedSKG (Xie et al., 2022), among others.

The pretrain-finetune paradigm In the context of multitask learning, cross-task generalisation was deemed an extremely challenging topic. This has changed with the relatively recent trend in which models are first *pretrained* with a general-purpose objective (language modelling, or masked language modelling) on large natural language corpora. The model is then further *finetuned* in a second stage, in which task-specific parameters are added that learn to execute different tasks using the representations learned in the pretraining stage. The popularisation of this *pretrain-finetune paradigm* has shifted

[7]Notably, as illustrated by the work of Weber et al. (2021), the definition of *task* can be taken liberally in this context, ranging from traditional notions of NLP tasks to considering subproblems of a single classic NLP task . For instance, while language modelling constitutes its own task, learning how to handle negative polarity items such as *any* or *ever* in a grammatically correct way can be considered a subtask of it.

thoughts on how to evaluate cross-task generalisation. Rather than evaluating how learning one task can benefit another, this paradigm instead gives a central role to the question of how well a model that has acquired some general knowledge about language during pretraining can be used to generalise to different kinds of tasks in a finetuning stage which involves task-specific parameters (e.g. Devlin et al., 2019; Howard and Ruder, 2018; Liu et al., 2019b; Peters et al., 2018). Interestingly, in the finetuning stage, performance on the tasks themselves is typically evaluated with random train–test splits, and thus generalisation within individual tasks is not necessarily considered.

Zero-shot and few-shot learning The focus of cross-task generalisation studies has more recently shifted even further, to scenarios which consider how well pretrained language models fare in different tasks without any task-specific parameters.[8] In the most extreme case, this implies evaluating a language model directly on a range of tasks without any further training. To do so, tasks are reformulated as text-completion problems, such that language models can be *prompted* directly with a question representing a specific task (*zero-shot learning*), potentially preceded by a few examples (*few-shot learning*) (Radford et al., 2019). The latter case, in which the intention is that models – without any parameter updates – 'learn' from the examples given in the context, is often referred to with the term *in-context learning*. Datasets for conducting tasks via prompting are typically created by adapting conventional multitask datasets, where prompting templates are (often manually) designed for each task (e.g. Mishra et al., 2022; Wang et al., 2022; Weller et al., 2020). Unfortunately, studies that investigate the relationship between the training and test data are rare, which leaves many open questions in this area. Where Brown et al. (2020) report that data leakage from training had a small impact on their results, other recent work suggests that the impressive capabilities of large language models on zero- or few-shot learning tasks can largely be attributed to the presence of similar or identical examples in the training corpus (Han and Tsvetkov, 2022; Razeghi et al., 2022). Moreover, models have been reported to be sensitive to exact task formulation (Jiang et al., 2020; Schick and Schütze, 2021) and even to the order of the examples given in the few-shot setting (Lu et al., 2022), to some extent contradicting the intuitive idea of task understanding – and thus being considered as evidence against models' generalisation ability.

In-context finetuning A different class of studies that considers task evaluation in the prompting setup are those that finetune a pretrained model with prompts from one set of tasks and then evaluate them on another set of tasks (e.g. Sanh et al., 2022; Wei et al., 2022; Zhong et al., 2021). Parallel to the term 'in-context learning', this scenario is often referred to with the term *in-context finetuning*. Here, the relationship between task performance and generalisation is clearer than in the zero- and few-shot learning setups. While also in this case the pretraining corpus is uncontrolled, at least the relationship between the finetuning training and test data can be monitored, and the performances on the test data with and without finetuning easily compared. Nevertheless, there are few studies that do so.

3.4 Generalisation across languages

A fourth type of generalisation is generalisation across languages, or cross-lingual generalisation. As described by Bender (2011), the availability of truly language-dependent NLP technologies would be very valuable from both a scientific and practical perspective. However, the field of NLP has been very biased towards models and technologies for English[9], and most of the recent breakthroughs rely on amounts of data that are simply not available for the vast majority of the world's languages. Cross-

[8] If the pretraining corpus is seen as a large collection of different uncontrolled tasks, this scenario is more similar to the original multitask learning scenario than the pretrain-finetune paradigm.

[9] To the point that, as pointed out in the same article from Bender (2011), studies that focus only on English do not even systematically report that this is the language that they are reporting results for.

lingual generalisation is thus extremely important to promote the inclusivity and democratisation of the field, as well as from a practical perspective.

Cross-lingual finetuning There are several ways in which cross-lingual generalisation can be evaluated. Most existing cross-lingual studies focus on the scenario where labelled data is available in a single language (typically English), and the model is evaluated in multiple languages. A common approach to address this problem is to finetune a multilingually pretrained language model on the English labelled data, and then transfer to the rest of the languages in a zero-shot fashion (e.g. Pires et al., 2019; Wu and Dredze, 2019).[10] For instance, Pires et al. (2019) show that Multilingual BERT (Devlin et al., 2019) finetuned on English generalises well even to languages with different scripts, but exhibits some systematic deficiencies that affect language pairs that have different word-order features, such as English and Japanese.

Multilingual learning A second way in which cross-lingual generalisation can be evaluated is by considering whether models trained on multiple languages at the same time (multilingual models) perform better than models trained on only one language. In multitask learning, approaches that are simultaneously trained on multiple tasks can be seen as an implicit evaluation of generalisation across tasks. Similarly, multilingual models trained on multiple languages can be seen as implicitly evaluating generalisation across languages. There is a large number of papers that investigates and proposes multilingual models, usually for language modelling or machine translation (e.g. Aharoni et al., 2019; Al-Shedivat and Parikh, 2019; Costa-jussà et al., 2022; Fan et al., 2021; Zhang et al., 2020). Most of these papers have as main aim to introduce improved models, and they are not motivated by generalisation questions. Some, however, do include explicit generalisation experiments in their setup. For instance, Zhou et al. (2018) investigate how generalisation depends on the amount of data added for different languages; whereas Aharoni et al. (2019) investigate how zero-shot generalisation changes depending on the number of different languages that a model is trained on.

Multilingual benchmarks As pointed out before, while the field of multilingual modelling is vast and associated with many interesting generalisation questions, papers in this area do not often focus explicitly on generalisation. We would, therefore, like to end this subsection by discussing the most important available multilingual benchmarks which can be used to evaluate cross-lingual generalisation. Multilingual benchmarks or datasets are created in a variety of ways. Several benchmarks are created by translating monolingual benchmarks into different languages, usually through a professional translation service (Artetxe et al., 2020; Conneau et al., 2018; Ebrahimi et al., 2022; FitzGerald et al., 2022; Lewis et al., 2020; Li et al., 2021a; Lin et al., 2021; Longpre et al., 2021; Mostafazadeh et al., 2016; Ponti et al., 2020; Williams et al., 2018; Xu et al., 2020; Yang et al., 2019; Zhang et al., 2019). Other multilingual benchmarks, instead, have been built by separately annotating each language via its native speakers (e.g. Adelani et al., 2021; Asai et al., 2021; Clark et al., 2020; Muller et al., 2021). Yet another way to construct multilingual benchmarks is to leverage existing resources that cover multiple languages. For instance, Wikipedia has been used as a resource to derive multilingual benchmarks (Botha et al., 2020; Liu et al., 2019a; Pan et al., 2017; Rahimi et al., 2019), and several multilingual summarisation datasets have been created by extracting article-summary pairs from online newspapers or how-to guides (e.g. Hasan et al., 2021; Ladhak et al., 2020; Nguyen and Daumé III, 2019; Scialom et al., 2020; Varab and Schluter, 2021). Various linguistic resources have also been exploited: for instance, the Universal Dependencies treebank (Nivre et al., 2020) has been used to evaluate cross-lingual part-of-speech tagging,

[10]Other approaches instead use machine translation to translate test sets into English and directly use an English model or to translate the training data into another language and finetune a multilingual model on the augmented data. As this setup does not focus on generalisation per se, but rather depends on the quality of the translation model, we will not further discuss it.

and multilingual WordNet and Wiktionary have been used to build XL-WiC (Raganato et al., 2020), an extension of WiC (Pilehvar and Camacho-Collados, 2019) that reformulates word sense disambiguation in 12 languages as a binary classification task. Finally, in the same spirit of GLUE and SuperGLUE for English, there are also several aggregated benchmarks that include selected sets of benchmarks previously proposed by others (e.g. Hu et al., 2020; Liang et al., 2020; Ruder et al., 2021; Wang et al., 2022), which allow for evaluating cross-task and cross-language generalisation simultaneously.

3.5 Generalisation across domains

The next category we include considers a type of generalisation that is often required in naturally occurring scenarios (more so than the types discussed so far) and is thus very important in practice: generalisation across different domains. As examples of the practical relevance of cross-domain generalisation, consider, for instance, a sentiment analysis model trained to classify the sentiment of reviews for certain products which then needs to generalise to newly commercialised products, necessarily not represented in its training data (Ryu et al., 2018; Tan et al., 2019); a model trained on data collected from one demographic which is then asked to generalise to the entire population (Blodgett et al., 2016); or a machine translation model trained on canonical text and then expected to generalise noisy data from an internet platform (Blodgett et al., 2017; Michel and Neubig, 2018) or to data from a different real-world domain (Malinin et al., 2021). While there is not a precise definition of what constitutes a domain, different domains broadly refer to collections of texts exhibiting different topical and/or stylistic properties, such as different genres or formality levels. Again, examples help us clarify this definition. MultiNLI (Williams et al., 2018), for instance, collects training corpora from five different genres (e.g. fiction and telephone conversations) and includes both an in-domain evaluation set with corpora from those five genres, as well as an out-of-domain evaluation set with corpora from five more sources (e.g. face-to-face conversations and the 9/11 public report). Blodgett et al. (2016) consider how language identification tools trained on Standard English generalise poorly to African-American English. Fried et al. (2019) compare how neural and non-neural constituency parsers generalise on out-of-domain treebanks (e.g. on a treebank of biomedical texts), whereas Artetxe et al. (2021) compare how sparse and dense language models generalise within and out of domain (on texts from ArXiv, Github, OpenSubtitles, among many other sources). Kamath et al. (2020) study the problem of selective question answering under domain shift, where the test distribution includes both in-domain and out-of-domain questions and the model must abstain from answering when not confident. Connected to this last type of study, there is a substantial body of work in out-of-domain *detection* (Hendrycks et al., 2020; Lane et al., 2007; Ryu et al., 2017, 2018; Tan et al., 2019).

Domain generalisation has often been studied in connection with domain adaptation, the problem of adapting an existing general model to a new domain (Daumé III, 2007). This has been a very active research area in machine translation (Axelrod et al., 2011; Bertoldi and Federico, 2009; Chu et al., 2017; Chu and Wang, 2018; Freitag and Al-Onaizan, 2016; Hu et al., 2019; Joty et al., 2015; Koehn and Schroeder, 2007; Luong and Manning, 2015; Wang et al., 2017a,b), with several standard datasets (Malinin et al., 2021; Michel and Neubig, 2018) and dedicated tracks in popular shared tasks like WMT (Bojar et al., 2019; Specia et al., 2020). In addition to machine translation, domain adaptation has also been studied in part-of-speech tagging (Blitzer et al., 2006), sentiment analysis (Blitzer et al., 2007) and language model pre-training (Gururangan et al., 2020), among others.

Finally, domain generalisation is closely related to temporal generalisation, where the training data is produced in a specific time period and the model is tested on data from a different time period, either in the future or in the past. This problem has been studied in an as yet limited range of tasks, including language modelling (Lazaridou et al., 2021), named entity recognition in social media (Derczynski et al., 2016; Fromreide et al., 2014; Rijhwani and Preotiuc-Pietro, 2020), named entity disambiguation (Agarwal et al., 2018), document classification (He et al., 2018; Huang and Paul, 2018, 2019) and sentiment analysis (Lukes and Søgaard, 2018).

3.6 Generalisation in the context of robustness

The last category of generalisation research we consider on the type axis considers how robust models are with respect to changes in their exact training data. We refer to such studies, that typically assess to what extent model performance is independent from the exact training data, with the term *robustness generalisation*. Studies of this kind usually focus on train–test shifts that stem from the data collection process. Different from most of the previous categories discussed in §3, such shifts are generally unintended and can be hard to spot. Existing research therefore focuses on characterising such scenarios and understanding their impact. This line of work is based on the idea that models should learn task solutions that abstract away over specific, often spurious correlations that may occur in the training data, i.e. models should learn the underlying generalising solution that humans associate with the task (e.g. Gururangan et al., 2018; McCoy et al., 2019; Talman and Chatzikyriakidis, 2019). Oftentimes, studies in this category intend to show that models do not generalise in the way we would expect them to, because the training data was in some very subtle manner not representative of the true target distribution. Robustness evaluation is very important from a practical perspective. If a model has a strong sensitivity to spurious patterns in the training data and is then tested on data collected with the same bias, this can result in overestimating its performance – either generally or on specific test cases – with potentially harmful consequences, for instance when a model does not generalise well to particular population demographics. Evaluating generalisation in the context of robustness can be driven by several different motivations. Some studies are motivated by more practical concerns, or are conducted to gain a better perspective on intrinsic task understanding, but robustness evaluation is also particularly important when the goal is to have fair and unbiased NLP models. Below, we discuss three common scenarios associated with robustness evaluation.

Annotation artefacts A scenario that often occurs in robustness studies is one where there are *annotation artefacts* in the training data, which may result in overestimation of a model's performance on a particular task. Artefacts occur particularly frequently when datasets are collected through crowdsourcing. Crowdsourced datasets often depend strongly on how exactly the annotation procedure was set up, with subtle artefacts as a consequence. For instance, annotators may naturally tend to minimise their cognitive effort, resorting to patterns that models learn to exploit. Popular NLI datasets like SNLI (Bowman et al., 2015a) and MultiNLI (Williams et al., 2018) have been found particularly susceptible to such artefacts. For instance, Gururangan et al. (2018) and Poliak et al. (2018) showed that a hypothesis-only baseline performs better than chance, due to its exploitation of spurious patterns in word choice and grammatical features (e.g. negation being indicative of the *contradiction* class). Talman and Chatzikyriakidis (2019) showed that NLI models do not generalise well across different datasets. Besides NLI, other tasks like question answering have also been reported to suffer from annotation artefacts (Jia and Liang, 2017; Kaushik and Lipton, 2018), even when such ertifacts were deliberately and consciously avoided during the annotation process (Elazar et al., 2021b). Finally, Lewis et al. (2021) showed that open-domain question answering datasets have a high overlap between train and test instances, and reveal that memorisation plays a bigger role in these benchmarks than previously assumed.

Standardised splits Another line of work questions the way we use data splits in general, and in particular the extent to which scores on standardised splits that stay static over time are reflective of a model's generalisation abilities. For instance, Gorman and Bedrick (2019) show that models perform much worse on random train–test splits than the reported state-of-the-art performances on a standardised split. Søgaard et al. (2021) go even further, and advocate for the use of heuristic and adversarial splits, where a model's capability for generalisation is challenged directly – for instance by putting all longer sentences in the test set, or by splitting the data to maximise the difference between train and test set.

Subpopulation bias A third scenario in which robustness and performance overestimation play a role is the case where certain demographics are under- or over-represented in the training data. As this may result in models that generalise poorly to specific demographic groups, it is a particularly harmful case of overestimation. For instance, Dixon et al. (2018) show that toxicity classifiers suffer from unintended bias, caused by certain identity terms being disproportionately represented in the training data (e.g. *"I am a gay man"* being assigned high toxicity scores because the word *"gay"* is often used in toxic comments). Similarly, Park et al. (2018) show that abusive language detection models exhibit gender bias, caused by imbalances in the training data. As a way to detect such imbalances and thus systematically avoid such cases of overestimation, Koh et al. (2021) propose to evaluate models by their worst-group accuracy, rather than the average accuracy across all demographic groups, in their CivilComments-Wilds dataset (a variant of the CivilCommons toxicity classification dataset released by Borkan et al., 2019).

4 Shift type: what kind of data shift is considered?

As we have seen in the previous two sections, tests to evaluate generalisation may differ in terms of their *motivation* and the *type* of generalisation that they target. What they share, instead, is that they all focus on cases in which there is a form of *shift* between the data a model is (pre)trained on and the data that is used for evaluation. In other words, for some datasets $(\mathcal{X}_1, \mathcal{Y}_1)$ and $(\mathcal{X}_2, \mathcal{Y}_2)$ considered in the experimental setup, it holds that $p(\mathbf{x}_1, \mathbf{y}_1) \neq p(\mathbf{x}_2, \mathbf{y}_2)$. In the third axis of our taxonomy, we discuss how to characterise shifts between the datasets used in a generalisation experiment. This axis adds a more statistical interpretation to our taxonomy and derives its importance from the fact that data shift plays an essential role in formally defining and understanding generalisation from a statistical perspective. On the data shift axis, graphically depicted in Figure 4, we consider three main types of shift which are well-attested in the literature: *covariate shift*, *label shift* and *full shift*. We further include two additional types of shift – *assumed shift* and *multiple shifts* – to account for studies that cannot be labelled with any of the three main shift types.

What are, precisely, data shifts? We formalise the differences between the test, training and potentially pretraining data involved in generalisation tests as shifts between the respective *data distributions*:

$$p(\mathbf{x}_{\text{tst}}, \mathbf{y}_{\text{tst}}) \qquad \texttt{test} \tag{1}$$

$$p(\mathbf{x}_{\text{tr}}, \mathbf{y}_{\text{tr}}) \qquad \texttt{training / finetuning / adaptation} \tag{2}$$

$$p(\mathbf{x}_{\text{ptr}}, \mathbf{y}_{\text{ptr}}) \qquad \texttt{pretraining} \tag{3}$$

By expressing these data distributions as the product of the probability of the input data $p(\mathbf{x})$ and the conditional probability of the output labels given the input $p(\mathbf{y}|\mathbf{x})$ –

$$p(\mathbf{x}_{\text{tr}}, \mathbf{y}_{\text{tr}}) = p(\mathbf{x}_{\text{tr}})\, p(\mathbf{y}_{\text{tr}}|\mathbf{x}_{\text{tr}}) \tag{4}$$

$$p(\mathbf{x}_{\text{tst}}, \mathbf{y}_{\text{tst}}) = p(\mathbf{x}_{\text{tst}})\, p(\mathbf{y}_{\text{tst}}|\mathbf{x}_{\text{tst}}) \tag{5}$$

we can define four main types of relations between any two data distributions.[11] One of these four types constitutes the case in which there is no shift in data distributions – i.e. both $p(\mathbf{x}_{\text{tr}}) = p(\mathbf{x}_{\text{tst}})$ and

[11]For clarity, we leave pretraining distributions aside and focus on train–test shifts, as this is the most intuitive setting. However, the shifts described in this section can be used to describe the relationship between any two data distributions involved in a modelling pipeline.

$p(\mathbf{y}_{\mathrm{tr}}|\mathbf{x}_{\mathrm{tr}}) = p(\mathbf{y}_{\mathrm{tst}}|\mathbf{x}_{\mathrm{tst}})$. This matches the i.i.d. evaluation setup traditionally used in machine learning. As discussed earlier, this type of evaluation, also referred to as *within-distribution* generalisation, has frequently been reported not to be indicative of good performance for the more complex forms of generalisation that we often desire from our models. We will not further discuss it here, but instead focus on the other three cases, commonly referred to as *out-of-distribution* (o.o.d.) evaluation.

Covariate shift The most commonly considered data distribution shift in o.o.d. generalisation research is one where $p(\mathbf{x}_{\mathrm{tst}}) \neq p(\mathbf{x}_{\mathrm{tr}})$ but $p(\mathbf{y}_{\mathrm{tst}}|\mathbf{x}_{\mathrm{tst}}) = p(\mathbf{y}_{\mathrm{tr}}|\mathbf{x}_{\mathrm{tr}})$. In this scenario, often referred to as *covariate shift* (Moreno-Torres et al., 2012; Storkey, 2009), the distribution of the input data $p(\mathbf{x})$ changes, but the conditional probability of the labels given the input – which describes the task – remains the same. Under this type of shift, one can evaluate if a model has learned the underlying task distribution while only being exposed to $p(\mathbf{x}_{\mathrm{tr}}, \mathbf{y}_{\mathrm{tr}})$. In NLP, covariate shift is a very common shift to evaluate in generalisation research. For example, challenge test sets such as HANS (McCoy et al., 2019), PAWS (Yang et al., 2019), or the COGS (Kim and Linzen, 2020) test set contain deliberately unusual, out-of-distribution examples, selected or generated to violate invalid heuristics in assigning labels to data samples. Less deliberate cases of covariate shift are evaluated in out-of-domain detection or robustness evaluation studies, such as those conducted by Ryu et al. (2018) and Tan et al. (2019) on real-world datasets. Tan et al. (2019), for instance, assume that the process by which the sentiment of a sentence is to be computed does not change, but the data that this process needs to be applied to does. Of the three o.o.d. shifts we discuss in this section, covariate shift is more easily addressed without performing additional training or pre- or post-processing than the other two shift types. As we will see in the next paragraphs, a common approach to address other, more complex shifts, is to turn them into covariate shifts.

Label shift The second type of shift corresponds to the case in which the focus is not on differences between the input distributions, $p(\mathbf{x}_{\mathrm{tst}}) = p(\mathbf{x}_{\mathrm{tr}})$, but instead in the conditional distributions of the labels/output: $p(\mathbf{y}_{\mathrm{tst}}|\mathbf{x}_{\mathrm{tst}}) \neq p(\mathbf{y}_{\mathrm{tr}}|\mathbf{x}_{\mathrm{tr}})$. We refer to this case as *label shift* but it is also known as *concept shift* (Moreno-Torres et al., 2012). Label shift can happen within the same task when there is a change of domain – e.g. the phrase *'it doesn't run'* can lead to different sentiment labels depending on whether it appears in a review for software or one for mascara; when there are inter-annotator disagreements; or when there is a temporal shift in the data (see §3.5). Another common case of label shift is a change in task (as in §3.3), where the meaning of the labels themselves changes as well. For example, the same sentence may need to be binarily classified for sentiment in some cases and for toxicity in others. In even more extreme cases, the labels themselves might change, for example when shifting from language modelling (where the set of labels is the language vocabulary) to POS-tagging. In NLP studies, label shift is often seen as an obstacle that needs to be overcome rather than as a setting in which models are directly evaluated: if the same example has contradictory labels in training and test data, it is unclear what decision at test time should be considered good generalising behaviour.

In practice, there are two main ways in which label shift is typically addressed. The first is to add an additional adaptation or finetuning stage, in which a model is updated to represent the shift that occurred (e.g. Biesialska et al., 2020; Sun et al., 2020), or new parameters are added to represent newly introduced labels (Devlin et al., 2019; Howard and Ruder, 2018; Peters et al., 2018, i.a.). In that scenario, there is a label shift between the pretraining and finetuning training data, but not between the finetuning training and testing data. The level at which generalisation is (somewhat implicitly) evaluated in that case, is the pretraining level: does my pretrained model adapt well to different conditional label distributions when further trained? The second way to address label shift is to augment the input data with domain or task indicators (e.g. Brown et al., 2020; Raffel et al., 2020). We saw before that the phrase *'it doesn't run'* can be both positive and negative, depending on what it describes. Without further information, it is impossible for a model to infer the correct meaning. By adding indicators that specify the domain (`review for mascara:...`, `review for software:...`), the problem is converted into a

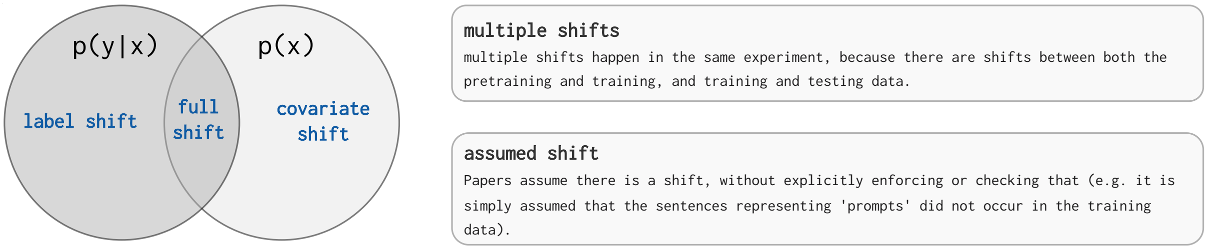


Figure 4: Types of data distribution shifts that can occur on the shift type axis of our taxonomy.

covariate shift (or potentially even no shift, if both indicators are represented in the two distributions at hand), which then can be solved by correctly generalising. Something similar happens in the case where a task is transformed into a question in a prompting setup: by adding a prompt that describes what needs to be done with the input, label shifts caused by a change of task are turned into a different type of shift that can be solved without further finetuning (see, e.g. Bach et al., 2022; Brown et al., 2020; Schick and Schütze, 2021).

Full shift The most extreme type of shift corresponds to the case in which both $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$ change simultaneously: $p(\mathbf{x}_{\text{tst}}) \neq p(\mathbf{x}_{\text{tr}})$ and $p(\mathbf{y}_{\text{tst}}|\mathbf{x}_{\text{tst}}) \neq p(\mathbf{y}_{\text{tr}}|\mathbf{x}_{\text{tr}})$. We refer to this case with the term *full shift*. Full shifts may occur in language modelling tasks, where changes in the $p(x)$ directly translate into changes in $p(y|x)$[12], or when adapting to new language pairs in multi-lingual experiments (e.g. Costa-jussà et al., 2022; Kodner et al., 2022). Another case of full shift is the one in which entirely different types of data are used either for pretraining (e.g. Papadimitriou and Jurafsky, 2020, who test if pretraining on music impacts learning language afterwards) or for evaluation (e.g. De Varda and Zamparelli, 2022, who evaluate generalisation to different languages). Oftentimes, covariate shifts might inadvertently also cause label shifts, for instance when the textual domain changes in a sequence-classification task. In our characterisation, however, if the underlying task stays the same, we will assume that the (more controlled) covariate shift is the one that is investigated, unless specified otherwise. Contrary to label shifts, full shifts can, in some cases, be addressed without retraining, because they do not necessarily imply that the same input x is assigned a different label at test time. However, similar to label shifts, also full shifts are often turned into different types of shifts that can be more easily addressed.

Multiple shifts In this section, we have considered three different data distributions and the types of shifts that can occur between any pair of such data distributions. Some studies, however, consider shifts between multiple distributions at the same time. For instance, Li et al. (2022) investigate how different types of pretraining architectures generalise to o.o.d. splits in a finetuning stage; and Wang et al. (2021) investigate which pretraining method performs better cross-domain generalisation in a second training stage. In our taxonomy, we label such cases *multiple shifts*, and – at least in the current version – we do not distinguish between different configurations of multiple shifts (e.g. label+covariate, or covariate+covariate). We will discuss multiple shifts further in §6.

4.1 On detecting shift type

We conclude this section by pointing out that while from a formal perspective the shifts that we describe are well-defined, they may be difficult to tell apart in practice because the base distributions by which

[12] An exception is the case in which a test consists of predicting only one word, such as, for instance, in a subject-verb agreement task. In that case, the predicted word is not ("autoregressively") part of the input of another prediction, and thus it does not automatically constitute a change in $p(y|x)$.

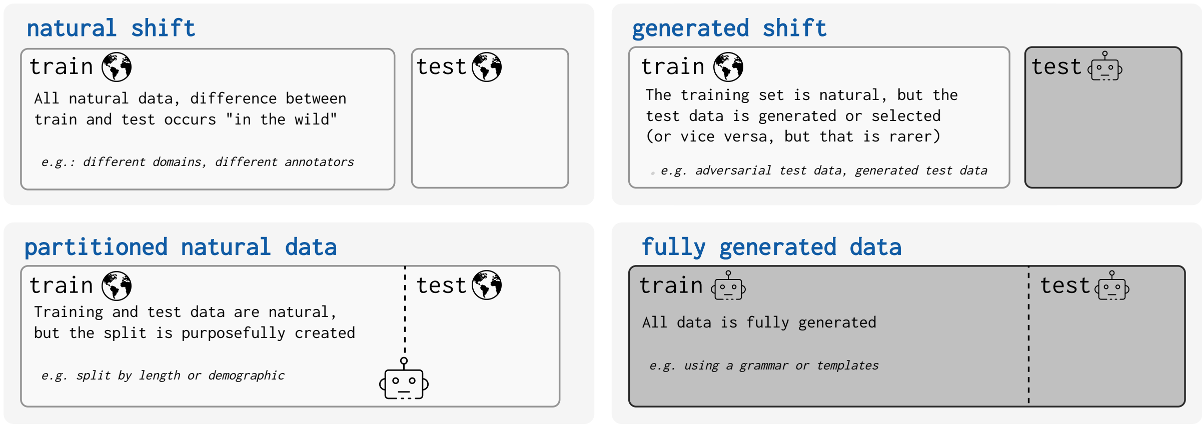


Figure 5: Different sources of shifts, with indications of what data is fully natural, indicated with a small globe, and data that is generated, indicated with a robot icon.

natural languages are 'generated' are rarely fully known. As a consequence, it is often not straightforward to determine what the relationship between two different datasets is. While in some cases there is nevertheless little discussion on the type of shift that occurs between two datasets, in other cases, it might be unclear if there is an actual shift, or what its nature is. When classifying shifts in our review, we will focus on cases where authors (i) explicitly consider the relationship between the data distributions they use in their experiments and (ii) the assumptions they make about this relationship are either well-grounded in the literature (e.g. it is commonly assumed that switching between domains constitutes a covariate shift) or empirically verified. Nevertheless, we identify numerous studies that claim to be about generalisation where such considerations are absent: it is *assumed* that there is a shift between train and test data, but this is not verified or grounded in previous research. Sometimes, the assumed shift is not explicitly checked because it is considered plausible given general (linguistic) knowledge about language. Consider, for instance, how Lakretz et al. (2021b), as discussed earlier in §3.2, study sentences with usually deep levels of recursion. Other times, the relationship between training and test data is not investigated because the researchers do not have access to the training data. The BigBench benchmark (Srivastava et al., 2022), for instance, contains several tasks that might measure generalisation, but the training datasets of the models investigated are not in the public domain. Yet in other cases, the training data is available to the authors of the paper, but simply no extensive analysis is presented (e.g. Brown et al., 2020; Chowdhery et al., 2022). In our survey, we also consider this entire body of work, which we mark *assumed shift*.

5 Shift source: how are the train and test data produced?

In the previous section, we discussed what types of shifts may occur in generalisation tests. We now focus on a related relevant dimension, that expresses how those shifts originated: our fourth axis, graphically shown in Figure 5, indicates the *source* of the differences occurring between the pretraining, training and test data distributions. The source of the data shift determines how much control the experimenter has over the training and testing data and, consequently, what kind of conclusions can be drawn from an experiment. Using fully generated data, for example, provides full control and allows to test very specific aspects in isolation, but might not be suitable to draw conclusions about a model's behaviour when it is exposed to a natural dataset. We distinguish four different sources of shifts: (i) *naturally occurring shifts*, shifts occurring naturally between different corpora; (ii) *splits of natural corpora*, in which the data distributions involved are all natural corpora, but they are artificially partitioned along a specific dimension; (iii) *generated shifts*, where the training data is natural, but the test data is designed with a specific distribution shift in mind;[13] and (iv) *fully generated datasets*, where all data involved is generated.

To formalise the description of these different sources of shift, we consider the unobserved *base distribution* which describes all data considered in an experiment:

$$p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}}, \boldsymbol{\tau}) \qquad\qquad \texttt{base} \tag{6}$$

The variable $\boldsymbol{\tau}$ represents a *data property of interest*, with respect to which a specific generalisation ability is tested. This can be an observable property of the data (e.g. the length of an input sentence), an unobservable property (e.g. the timestamp that defines when a data point was produced), or even a property relative to the model (architecture) under investigation (e.g. $\boldsymbol{\tau}$ could represent how quickly a data point was learned in relation to overall model convergence). The base distribution over $\mathbf{x}$, $\mathbf{y}$ and $\boldsymbol{\tau}$ can be used to define different partition schemes, which can be adopted in generalisation experiments. Formally, such a partitioning scheme is a rule $f : \mathcal{T} \rightarrow \{\texttt{true, false}\}$ that discriminates data points according to a property $\boldsymbol{\tau} \in \mathcal{T}$. To investigate how a partitioning scheme impacts model behaviour, the pretraining, training and test distributions can be defined as:

$$p(\mathbf{x}_{\text{ptr}}, \mathbf{y}_{\text{ptr}}) = p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} \,|\, f_{\texttt{pretrain}}(\boldsymbol{\tau}) = \texttt{true}) \tag{7}$$
$$p(\mathbf{x}_{\text{tr}}, \mathbf{y}_{\text{tr}}) = p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} \,|\, f_{\texttt{train}}(\boldsymbol{\tau}) = \texttt{true}) \tag{8}$$
$$p(\mathbf{x}_{\text{tst}}, \mathbf{y}_{\text{tst}}) = p(\mathbf{x}_{\text{base}}, \mathbf{y}_{\text{base}} \,|\, f_{\texttt{test}}(\boldsymbol{\tau}) = \texttt{true}) \tag{9}$$

Using these data descriptions, we can now discuss four different sources of shifts.

Naturally occurring shifts The first scenario we consider is the one in which shifts naturally occur between different corpora. In such cases, the variable $\boldsymbol{\tau}$ refers to properties that naturally differ between collected datasets. What characterises this type of shift source, is that both the data partitions of interest are naturally occurring corpora, to which no systematic operations are applied: for the purposes of a generalisation test, experimenters have no direct control over the partitioning scheme $f(\boldsymbol{\tau})$. Examples of naturally occurring shifts emerge from splits containing data from different annotators (Geva et al., 2019), sources or domains (e.g. Artetxe et al., 2021; Talman and Chatzikyriakidis, 2019), data sampled from different populations (e.g Dixon et al., 2018; Talat et al., 2018) data from different points in time (e.g. Lazaridou et al., 2021), or separately collected corpora targeting the same task, such as MNLI (Williams et al., 2018) and WNLI (Wang et al., 2018). In this category, we also include cross-task and cross-lingual generalisation studies in which all corpora involved are natural corpora (e.g. FitzGerald et al., 2022; Mishra et al., 2022).

Splits of natural corpora A slightly less natural setup is the one in which a natural corpus is considered, but it is artificially split along specific dimensions. The primary difference with the previous category is that the variable $\boldsymbol{\tau}$ refers to data properties along which data would not naturally be split, such as the length or complexity of a sample. The experimenters have thus no control over the data itself, but they do control the partitioning scheme $f(\boldsymbol{\tau})$. Raunak et al. (2020), for instance, split naturally occurring machine translation corpora such that longer sentences occur in the test data, and Weber et al. (2021) split a language modelling corpus such that the training data does not contain specific types of negative polarity item licensers. Other examples of natural data splits could be splits that maximise compound divergence (Keysers et al., 2019) to investigate compositionality.[14]

[13] Or, more rarely, the other way around.

[14] Keysers et al. (2019) themselves do not apply this split to fully natural data, their corpus is fully generated using templates.

Generated shifts The third category on our source of shift axis concerns the case in which one data partition (usually the *training* set) is a fully natural corpus, but the other partition is designed with specific properties in mind, to address a generalisation aspect of interest. Data in the constructed partition may avoid or contain specific (syntactic) patterns (Bhargava et al., 2021; Cui et al., 2022), violate heuristics about gender (Dayanik and Padó, 2021; Libovický et al., 2022), or include unusually long or complex sequences (Lakretz et al., 2021a; Raunak et al., 2019). As an example of this shift source, Dankers et al. (2022) investigate compositionality in MT models trained on fully natural corpora by constructing test data that addresses compositional generalisation given the specific properties of the training corpus. For NLI, McCoy et al. (2019) design a test set that cannot be solved with models that rely on specific heuristics. Fancellu et al. (2017) create a test set for which the select sentences with negation scopes that are not delimited by punctuation. Another category of studies that fit into this type are those with *adversarial* test sets, generated either by humans (Kiela et al., 2021) or automatically using a specific model (e.g. Sakaguchi et al., 2021; Zellers et al., 2018). In the examples above, all of the constructed data occurs in the test data; note that the opposite – where instead the *training data* is synthetic or generated and the test data natural – is also possible, yet less common (e.g. Papadimitriou and Jurafsky, 2020).

Fully generated The last category we consider are splits that use only generated data, which sometimes may even be fully synthetic. Generating data is often the most precise way of measuring specific aspects of generalisation, as experimenters have direct control over both the base distribution and the partitioning scheme. Sometimes the data involved is entirely synthetic (e.g. Hupkes et al., 2020; Lake and Baroni, 2018), other times it is templated natural language or a narrow selection of an actual natural language corpus (e.g Keysers et al., 2019; Kim and Linzen, 2020). Generated splits can vary in several different dimensions. Sometimes, $\boldsymbol{\tau}$ is a simple observable data property. For instance, Hupkes et al. (2020) split their corpus based on the presence of particular function pairs $\mathcal{P}$, implicitly setting $\boldsymbol{\tau} = \mathcal{P} \in x$. In some cases, $\boldsymbol{\tau}$ may also be defined relative to the $\boldsymbol{\tau}$ of other examples, and can only be computed globally, such as in the case of *maximum compound divergence* splitting (Keysers et al., 2019).

6 Locus of shift: between which data distributions does the shift occur?

In the previous sections, we discussed high-level motivations for studying generalisation in NLP models, types of generalisation that have been frequently evaluated in the literature, kinds of data distribution shifts used for generalisation tests, and the possible sources of those shifts. These four axes demonstrate the depth and breadth of generalisation evaluation research, and they also clearly illustrate that generalisation is evaluated in a wide range of different experimental setups. What we have not yet explicitly discussed is between which data distributions those shifts can occur: the *locus* of the shift. In our taxonomy, the shift locus forms the last piece of the puzzle, as it determines what part of the modelling pipeline is investigated and, with that, what kind of generalisation questions can be asked. For instance, shifts between pretraining and training distributions allow the experimenter to investigate if a particular pretraining procedure is successful, whereas train–test shifts can be used to evaluate a model instance or a training procedure. We consider shifts between all stages in the contemporary modelling pipeline – pretraining, training and testing, as well as studies that consider shifts between multiple stages at the same time, as expressed by the data distributions that we have considered in §4 (for a graphical representation, we refer to Figure 6).

Given these distributions, there exist five possible loci of shifts: shifts only between the (finetune) *training and the test data*, shifts only between the *pretraining and the training data*, shifts only between the *pretraining and the test data*, and shifts between *all data distributions*. Because they often reflect different types of experiments, we separate shifts between train and test data without pretraining from

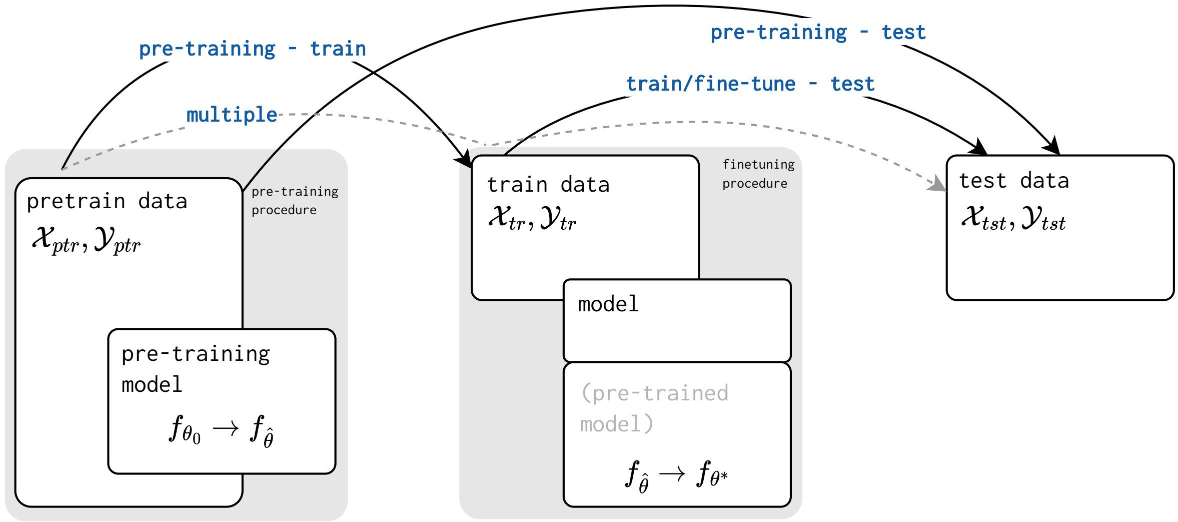


Figure 6: Different loci of splits, and what parts of the modelling pipeline they may investigate generalisation for.

shifts between finetuning train and test data. We describe the four loci of shift and how they interact with different components of the modelling pipeline with the aid of three *modelling distributions*. These modelling distributions correspond to the different stages in contemporary machine learning pipelines – testing a model, training it, and potentially pretraining it:

$$p(\mathcal{Y}_{\text{tst}} \mid \mathcal{X}_{\text{tst}}, \boldsymbol{\theta}^*) \qquad \texttt{model} \qquad (10)$$

$$p(\boldsymbol{\theta}^* \mid \mathcal{X}_{\text{tr}}, \mathcal{Y}_{\text{tr}}, \phi_{tr}, \hat{\boldsymbol{\theta}}) \qquad \texttt{training/finetuning/adaptation} \qquad (11)$$

$$p(\hat{\boldsymbol{\theta}} \mid \mathcal{X}_{\text{ptr}}, \mathcal{Y}_{\text{ptr}}, \phi_{pr}, \boldsymbol{\theta}_0) \qquad \texttt{pretraining} \qquad (12)$$

In these equations, ϕ broadly denotes training and pretraining hyperparameters, $\boldsymbol{\theta}$ refers to model parameters, and $\mathcal{X}, \mathcal{Y}$ indicate sets of inputs ($\mathbf{x}$) and their corresponding output ($\mathbf{y}$). In short, Equation 10 defines a model instance, which specifies the probability distribution over the target test labels $\mathcal{Y}_{\text{tst}}$, given the model's parameters $\boldsymbol{\theta}^*$ and a set of test inputs $\mathcal{X}_{\text{tst}}$. Equation 11, instead, defines a training procedure, specifying a probability distribution over model parameters $\boldsymbol{\theta}^* \in \mathbb{R}^d$ given a training dataset $\mathcal{X}_{\text{tr}}$, $\mathcal{Y}_{\text{tr}}$, a set of training hyperparameters ϕ_{tr}, and a (potentially pretrained) model initialisation $\hat{\boldsymbol{\theta}}$. Lastly, Equation 12 defines a pretraining procedure, specifying a conditional probability over the set of parameters $\hat{\boldsymbol{\theta}}$, given a pretraining dataset, a set of pretraining hyperparameters ϕ_{pr}, and a model initialisation.[15] Between which of these stages a shift occurs impacts which of these modelling distributions can be evaluated. We discuss the different potential loci of shifts below.

The train–test locus Probably the most commonly occurring locus of shift in generalisation experiments is the one between train and test data. This locus occurs in the classic setup where a model is trained on some training data and then directly evaluated on a shifted (out-of-distribution) test partition. Studies with the train–test locus can assess two different parts of the modelling pipeline. In some cases, researchers investigate the generalisation abilities of a *model instance* (i.e. a set of parameters $\boldsymbol{\theta}^*$, as described in Equation 10). Studies of this type therefore report the evaluation of a single model instance – typically made available by others – without considering how exactly it was trained, and how that impacted the model's generalisation behaviour. For example, a surge of studies considered the behaviour

[15]Note that this formalisation generalises to the *training from scratch* paradigm when $\mathcal{X}_{\text{ptr}}, \mathcal{Y}_{\text{ptr}} = \emptyset, \emptyset$, and to the *in-context-learning* setup when $\mathcal{X}_{\text{tr}}, \mathcal{Y}_{\text{tr}} = \emptyset, \emptyset$.

of the pretrained language model made available by Gulordava et al. (2018), to investigate how it generalised to, for instance, different syntactic constructions (e.g. Lakretz et al., 2019).[16] Alternatively, researchers might evaluate one or more training procedures, by considering if the *training distribution* results in model instances that generalise well – for example, to study how generalisation compares between dense and sparse models or how that changes with the scale of the input data (e.g. Artetxe et al., 2021; Rae et al., 2021), or how different architectures behave on a compositional generalisation test (Mul and Zuidema, 2019; Saxton et al., 2019). While also this case requires evaluating model instances, the focus of the evaluation is not on one particular model instance, but rather on the procedure that generated multiple model instances.

The finetune train–test locus The second potential locus of shift bears similarities to the first one but instead considers data shifts between the train and test data during finetuning, considering a model that has already gone through an earlier stage of training. This locus occurs when a model is evaluated on a finetuning test set that contains a shift with respect to the finetuning training data. An example of this category would be a test that investigates how well one pretrained model generalises with respect to an o.o.d. finetuning train–test split (Damonte and Monti, 2021; Kavumba et al., 2022; Ludwig et al., 2022). The parts of the modelling pipeline that studies with a finetune train–test locus can evaluate are the same as studies with a train–test locus, although studies that investigate the generalisation abilities of a single finetuned model instance are rare. More frequently, research with this locus focuses on the finetuning procedure, by considering if it results in finetuned model instances that generalise well on the finetune test set. Note that studies evaluating o.o.d. splits during finetuning, often also include a comparison between different pretraining procedures (e.g. they investigate whether BERT or RoBERTa generalises better to an o.o.d. finetuning test set, or compare how BERT models trained on different corpora behave during finetuning). Such studies (usually) investigate both a shift from the pretraining to the finetuning training data (typically a label shift), as well as a shift in the finetuning stage, and we will mark them as having *multiple loci*, as will be further discussed in the last paragraph of this section.

The pretrain-train locus A third potential locus of shift is between the pretraining and training corpus. Experiments with this locus evaluate whether a particular pretraining procedure, as described in Equation 12, results in models (parameter sets $\hat{\boldsymbol{\theta}}$) that are useful when further trained on different tasks or domains. For instance, Artetxe et al. (2021) investigate which pretraining procedure shows the best downstream generalisation in a number of different tasks, Tian et al. (2021) investigate how well pretrained models generalise to a newly proposed first-order-logic dataset, and Freitag and Al-Onaizan (2016) test how well a pretrained NMT model can adapt to different domains. Crucially, we classify studies as having a pretrain-train locus only when in their second training stage – which is required to have this locus – they use i.i.d. splits. If also the finetuning stage contains a shift, we say that the study has *multiple loci*.

The pretrain–test locus The fourth potential locus of shift is between pretraining and test data. This locus occurs when a pretrained model is not further updated but evaluated directly (i.e. $\mathcal{X}_{\mathrm{tr}}, \mathcal{Y}_{\mathrm{tr}} = \emptyset, \emptyset$) – as frequently happens in in-context learning setups (e.g. Lin et al., 2021; Zhang et al., 2022) – or when a pretrained model is finetuned on examples that are i.i.d. with respect to the pretraining data and then tested on out-of-distribution instances. The former case ($\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}$) is similar to studies with only one training stage in the train–test locus, but distinguishes itself by the nature of the (pre)training procedure, which typically has a general purpose objective, rather than being task-specific (e.g. a language modelling objective). Furthermore, while generalisation studies with a train–test locus almost always

[16]The investigation of model instances is, however, more common with the *pretrain-test* locus that we will discuss later in this section.

explicitly consider the relationship between training and test data, this is frequently not the case with pretrain–test studies in an in-context learning or finetuning setup: often, they do not explicitly consider the relationship between training and test data, but merely assume a shift occurs between those stages (e.g. Radford et al., 2019).

Multiple loci The last option on our locus axis is the *multiple loci* case, which we use for works that consider, in a single study, multiple shifts between different parts of the modelling pipeline. More explicitly, experiments of this type present shifts both between the pretraining and training data, as well as between the training and test data.[17] Multiple-loci experiments evaluate all stages of the modelling pipeline at once: they consider both how generalisable the models produced by the pretraining procedure are, as well as whether generalisation happens in the finetuning stage itself. For instance, some studies compare how well models with different pretraining procedures (e.g. BERT vs RoBERTa) generalise to o.o.d. splits during finetuning (e.g. Tu et al., 2020), others how different multilingual pretraining procedures perform cross-lingual task generalisation in a finetuning stage (e.g. FitzGerald et al., 2022; Hu et al., 2020; Yanaka et al., 2021). Because multiple-loci experiments necessarily also contain multiple shifts, we mark them as *multiple shifts* in the shift type axis. The nature of these shifts may not be the same: the shift from pretraining to training may be of any type, while the shift from training to test is often – but not necessarily – a less extreme covariate shift. In the current version of the taxonomy, we do not further distinguish these cases but collapse them into a single 'multiple shifts' category.

7 A review of existing generalisation research

In this paper, we have presented a taxonomy containing five categorical axes that can be used to characterise generalisation research. We now use our taxonomy to analyse a large amount of existing generalisation research and create a comprehensive map indicating which areas are covered and which are still unexplored. On our website[18], we present interactive ways to visualise our results and to retrieve relevant citations, which the reader can use to get a more in-depth view, to understand how their work fits in with the rest of the literature or which areas might be promising to address. We provide instructions for other researchers to contribute to the review, for instance by proposing to add new studies and studies we may have missed or by proposing corrections to studies that might have been misqualified on one of their axes values. In this section, we present our main findings.

7.1 Setup

We first briefly describe the procedures we used for the selection of the papers in our review and their annotation.

Paper selection An initial selection of manuscripts was made through a substantive preliminary literature review by the main authors of this paper. We then carried out a search through the ACL anthology. We started by retrieving all papers that have the (sub)words *generalisation*, *generalization*, *generalise* or *generalize* in their title or abstract. In Figure 8, we see that the number of papers with those keywords grew substantially over time, both in absolute and relative terms. We manually checked the abstracts and titles of the resulting papers to remove those that were not, in fact, addressing a generalisation question (for instance, because they proposed a generalisation of a *method*, or because they used random train–test splits). Furthermore, we restricted ourselves to papers with one modality. We then annotated

[17] We do not distinguish cases where the test data is shifted with respect to the pretraining data from cases where it is not, as the latter are very uncommon. It is, however, possible to set up an experiment where the pretraining and test data are drawn from the same distribution, for example to test whether a finetuning procedure results in catastrophic forgetting.

[18] `https://genbench.github.io/visualisations`

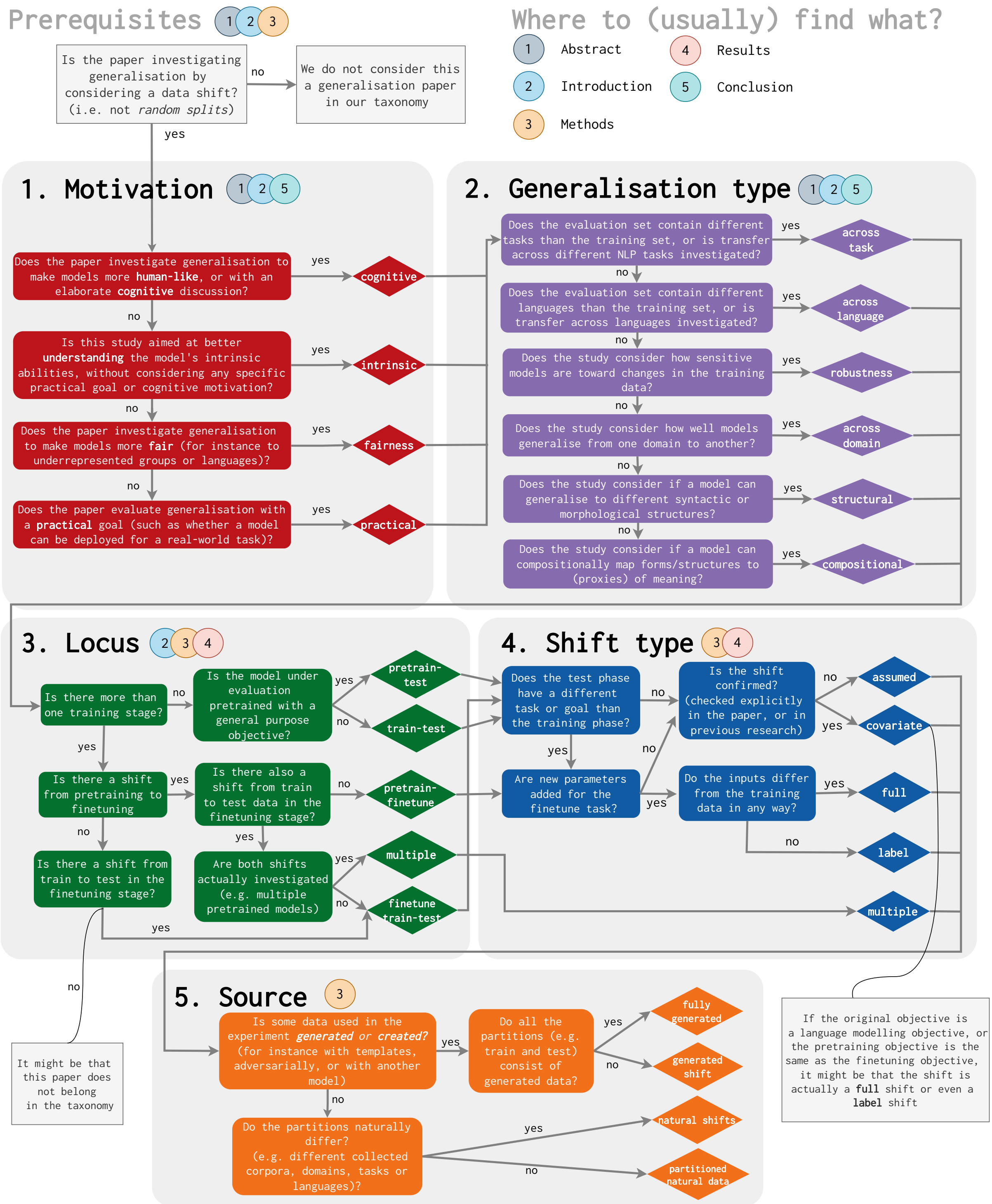
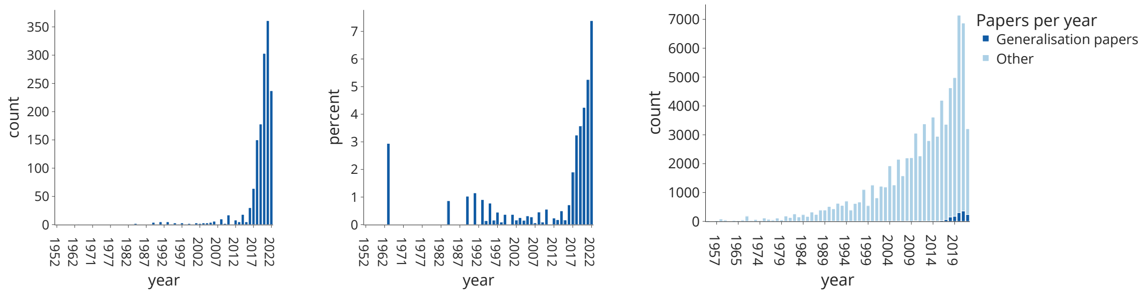


Figure 7: A graphical representation of our annotation process and an indication of where in a paper you might find the information required to complete the annotation. One paper can potentially contain multiple generalisation questions – e.g. both cross-domain and cross-task generalisation, or both generated shifts and splits using natural data. In that case, the diagram has to be walked through twice. Of course, the diagram is an aid that helps characterise papers but also simplifies the full taxonomy. On our website, we keep track of common questions that arise when using the diagram to characterise papers in an FAQ.



(a) (b) (c)

Figure 8: We selected papers from the ACL anthology that contain the (sub)words *generalisation*, *generalization*, *generalise* or *generalize* in their title or abstract. This figure shows how many of such papers exist per year, both absolutely (a) and percentually (b). In (c), we also show the total number of papers and generalisation papers published each year.

the resulting papers using the taxonomy presented in the previous sections. During the annotation process, we sometimes removed entries that upon further reading did not, in fact, contain generalisation experiments, and we duplicated entries that contained multiple experiments with different values on one of our axes. The findings presented in this section encompass in total 619 generalisation experiments, presented in 449 papers. The full list of papers can be found in the second bibliography at the end of this paper, as well as on our website[19]. While the conclusions in this – static – paper pertain only to this specific selection of papers, we intend to keep expanding the number of entries on our website with existing papers we missed or as new generalisation papers are published.

Annotation The annotation of all selected papers was done collectively by the authors of this article. Each paper was given five labels by a first annotator, one for every axis of our taxonomy, and these labels were then checked by a second annotator. Disagreements were discussed among the two annotators, and for unresolved cases, a third annotator was used. As a guide, we used the diagram presented in Figure 7. An FAQ with common questions that occurred while using this diagram, which intends to capture our taxonomy but is naturally a simplified version of it, can be found on our website. In addition to the taxonomy axes values, we also annotated which task(s) the studies considered. If a paper performed the same experiment with multiple different tasks, we label it *multiple tasks*, use the overarching category (e.g. *NLU*) when possible, or mark it as *multitask* if the purpose is to show that a paper can do those all at the same time. If a paper contained multiple studies with different values on the same axis – e.g. a paper considers both cross-domain and compositional generalisation or uses both natural shifts and synthetic data – we record those experiments separately.

7.2 Results

We now proceed to present the main conclusions drawn from our review, in particular focusing on overall trends for each axis (§7.2.1) and on how the different axes interact with each other (§7.2.2).

7.2.1 Overall trends on different axes

First, we discuss the overall occurrences of values on all axes, without taking into account interactions between them. We plot the (relative) occurrences of all values in Figure 9 and their development over time in Figure 10. Because the number of generalisation papers before 2018 included is very low (see

[19]https://genbench.github.io/references

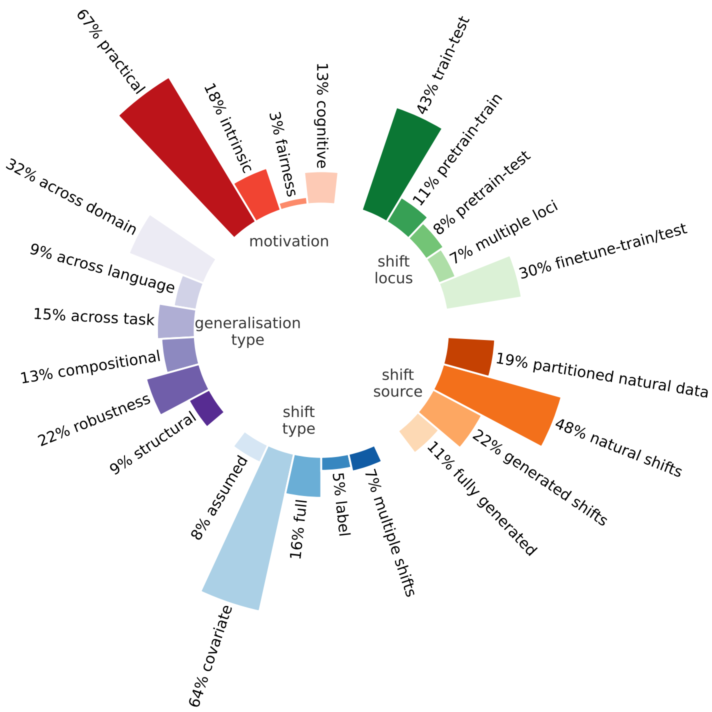
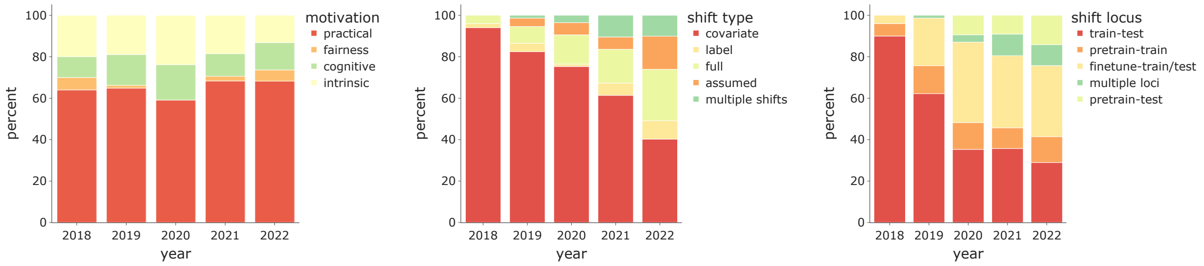


Figure 9: Summary plot displaying the relative occurrences of the categories available within the five different axes of our taxonomy (shown clockwise are the motivation, the generalisation type, the shift source, the shift type and the shift locus).

Figure 8a), we restricted the over-time plots to the last five years; all other statistics reported are computed over all papers.

Motivations As we can see in Figure 9 (top left), by far the most common motivation to test generalisation is the practical motivation. The intrinsic and cognitive motivations follow, whereas the studies in our review that consider generalisation from a fairness perspective make up only 3% of the total. We hypothesise that one of the reasons that this percentage is so low stems from the fact that our keywords search in the anthology was not optimal for detecting fairness studies, and we welcome researchers to suggest other generalisation studies with a fairness motivation for review. We will include them in an updated version of this paper. However, we also speculate that only relatively recently attention for the potential harmfulness of models trained on large, uncontrolled corpora is starting to grow and that fairness has simply not been studied as much in the context of generalisation yet. Due to the extremely low number of fairness studies in our review, it is not possible to observe a reliable growth of fairness papers in the last few years. In Figure 10a, we see that trends on the motivation axis have some small fluctuations over time but have been relatively stable over the past five years.

Generalisation type For generalisation types (Figure 9, left side), we find that cross-domain is the most frequent, making up more than 30% of all studies, followed by robustness, cross-task and compositional generalisation. Structural and cross-lingual generalisation are the least commonly investigated. As already mentioned in the respective section, studies looking at the understanding of syntactic and morphological structure typically focus more on whether models can capture structures at all, rather than on whether they generalise to new structures, which could be a potential explanation for the fact



(a) Motivation (b) Shift type (c) Shift locus

Figure 10: Trends from the past five years for three of the taxonomy's axes (motivation, shift type and shift locus), normalised by the total number of papers annotated per year.

that such studies are underrepresented. The underrepresentation of cross-lingual studies could, similar to studies with a fairness motivation, be partly explained by the fact that they might less frequently use the word generalisation in their title or abstract. However, we hypothesise that, at least in part, the low numbers are also reflective of the English-centric approach that is usually taken in NLP. As with fairness studies, we encourage researchers to suggest cross-lingual generalisation studies that we may have missed via our website so that we can determine better to what extent cross-lingual studies are, in fact, underrepresented.

Shift type Data shift types (Figure 9, bottom) are very unevenly distributed over their potential values: the vast majority of generalisation research considers covariate shift. Given the fact that covariate shift can occur between any two stages in the modelling pipeline, and label and full shift typically only occur between pretraining and finetuning, this is – to some extent – to be expected. Furthermore, covariate shift is more easily addressed by most current modelling techniques. More unexpected, perhaps, is the relatively high amount of *assumed* shifts, which correspond to studies that claim to test generalisation but do not explicitly consider how the test data relates to data used at various stages of model training. In Figure 10b, we see that the percentage of assumed shifts has increased over the past few years. We hypothesise that this trend, which is a step in the wrong direction in that it indicates less precision about what we evaluate rather than more, is predominantly caused by the use of increasingly large, general-purpose training corpora. Such large corpora, which are often also not in the public domain, make it very challenging to analyse the relationship between the training and testing data and, consequently, make it hard to determine what kind of conclusions can be drawn based on test accuracies. More promising, instead, is the fact that several studies consider *multiple shifts*, meaning that they assess generalisation throughout the entire modelling pipeline rather than only in one stage.

Shift source On the shift source axis (Figure 9, bottom right), we see that almost half of the reviewed generalisation studies consider naturally occurring shifts: natural corpora that are not deliberately split along a particular dimension. As we will see later, this type of data source is most prevalent in cross-task and cross-domain generalisation studies, for which such naturally different corpora are widely available. The next most frequent category is generated shifts, where one of the datasets involved is generated with a specific generalisation property in mind, and artificially partitioned natural data, describing settings in which all data is natural, but the way it is split between train and test is not. Fully generated datasets are less common, making up only 11% of the total number of studies.

Shift locus Lastly, for the locus axis (Figure 9, top right), we see that the majority of cases focuses on (finetune) train–test splits. Much fewer studies focus on shifts between pretraining and training or pretraining and testing. Similar to the previous axis, we observe that a comparatively small percentage of studies considers shifts in multiple stages of the modelling pipeline. We hypothesise that, at least in part, this might be driven by the larger amount of compute that is typically required for those scenarios. In Figure 10c, however, we also see an alternative explanation for the lower overall frequency of studies considering multiple loci and pretrain–test loci: the values populating Figure 9 are averaged over all years represented in our paper selection, but the multiple and pretrain–test loci became more popular only in the last few years.

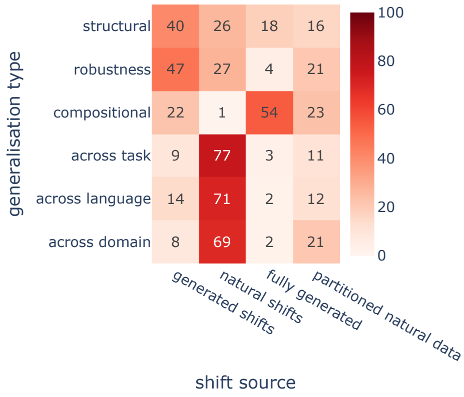
7.2.2 Interactions between axes

Next, we consider interactions between different axes. Are there any combinations of axes that occur together very often or combinations that are instead rare? We encourage the reader to view these interactions dynamically on our website. Here, we discuss a few trends.

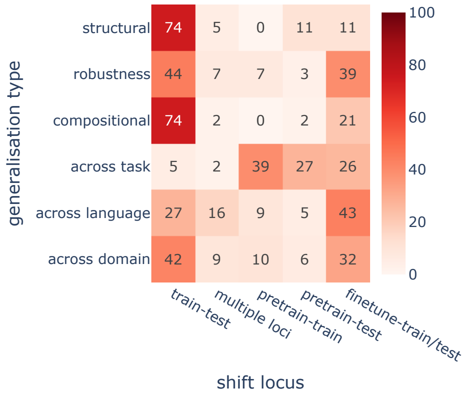
What data shift source is used for different generalisation types? In Figure 11a, we plot the frequency of each data source per generalisation type, normalised by the total number of times that generalisation type occurs (to make patterns comparable between generalisation types). From this plot, we can see that the type of data used is vastly different across different types of generalisation tests. Compositional generalisation, for instance, is predominantly tested with fully generated data, a data type that hardly occurs in research considering robustness, cross-lingual or cross-task generalisation. Those three types of generalisation are most frequently tested with naturally occurring shifts or, in some cases, with artificial splits of natural corpora. Structural generalisation, on the other hand, is the only generalisation type that appears to be tested across all different data types. As far as we know, there are very few studies that directly compare results between different sources of shift – for instance, to investigate to what extent results on generated shifts or fully generated data are indicative of performances on natural corpora.[20] Such studies could provide insight into how choices in the experimental design impact the conclusions that are drawn from the experiment, and we believe that they are an important direction for future work.

For which loci of shift are different generalisation types studied? Another interesting question to ask is for which locus different generalisation types are considered. In Figure 11b, we see that of all the generalisation types, only cross-task generalisation is frequently investigated in the pretrain-train and pretrain–test stages. For all other types of generalisation, the vast majority of tests are conducted in the train–test or finetune-train/test stage. In some cases, these differences are to be expected: as general-purpose pretrained models are usually trained on very large, relatively uncontrolled corpora, investigating how they generalise to a different domain without further finetuning is typically not possible, and neither is evaluating their robustness, which typically also requires more detailed knowledge of the training data. The statistics also confirm the absence of studies that consider compositional generalisation from pretraining to finetuning, or even from pretraining to training, which as we previously reported (§3.1) is philosophically and theoretically challenging in such setups. A final observation is the relative under-representation of studies with multiple loci across all generalisation types, especially given the large number of studies that consider generalisation in the finetuning stage or the pretrain-training stage. Those studies have used both a pretraining and finetuning stage but considered generalisation in only one of those. We hope to see this trend changing in the future, with more studies considering generalisation in the entire modelling pipeline, rather than only in a specific part of it.

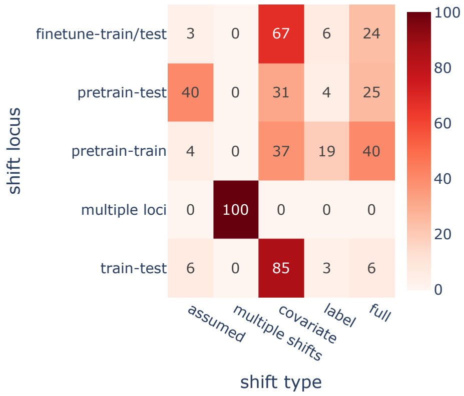
[20] An example of such a study would be the work of Chaabouni et al. (2021), who investigate whether performance improvements on SCAN transfer to machine translation models trained on natural data.



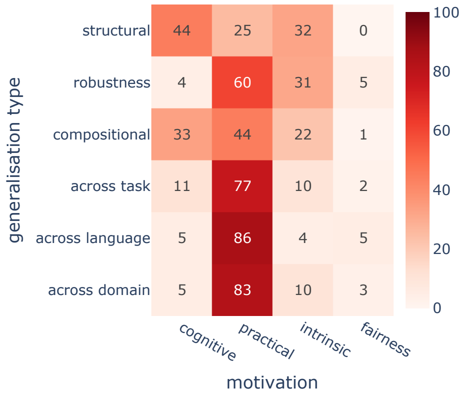
(a) Data source per generalisation type



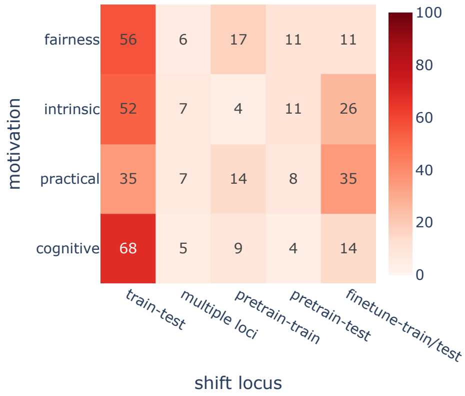
(b) Shift locus per generalisation type



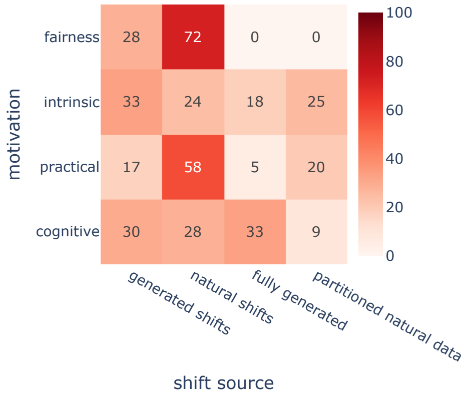
(c) Shift type per locus



(d) Motivation per generalisation type



(e) Locus per motivation



(f) Shift source per generalisation type

Figure 11: Heatmaps of interactions between axes. The maps are normalised by the total row value. This facilitates the comparison of patterns between rows but renders columns incomparable. We welcome readers who would like to see different normalisations or readers that are curious about interactions between other axes to have a look at our website, where they can generate other plots based on the same underlying data.

Which types of data shifts occur across different loci? Another interaction we would like to discuss is the one between the shift locus and the type of data shift. We plot this interaction in Figure 11c. A notable observation is that assumed shifts mostly occur in the pretrain–test locus, which confirms our hypothesis put forward earlier when discussing frequencies on the shift type axis – that assumed shifts are likely caused by the use of increasingly large, general-purpose training corpora. When such pretrained models are further finetuned, they often consider either a shift between pretraining and finetuning where new labels are introduced, or a covariate shift in the finetuning stage and, as such, do not require an in-depth understanding of the pretraining corpus.[21] When such models are directly evaluated, however, the only shift that can be considered is the one between the very large pretraining corpus and the test corpus. This trend points to a substantial challenge when it comes to evaluating generalisation for models with limited knowledge about their pretraining.

How does motivation drive generalisation research? The last pattern we would like to discuss is the relationship between the motivation behind a study and the other axes, focusing in particular on generalisation type, shift locus and shift source, as shown in Figure 11d-11f. Considering first the relationship between motivation and generalisation type (Figure 11d), we see that cross-domain, robustness, cross-task and cross-lingual generalisation are predominantly motivated by practical considerations. Robustness generalisation studies are also frequently motivated by the interest in understanding how models work (the *intrinsic* motivation). When looking at compositional and structural generalisation studies, we see that both are frequently driven by cognitive motivations – which is to be expected given the importance of these concepts in human cognition and intelligence. The motivation given most frequently for compositional generalisation, however, is a practical one. While in human learning, compositionality is indeed often associated with important practical properties – speed of learning, quick generalisation – as far as we know, there is little empirical evidence that compositional models actually perform better for natural language tasks. A similar apparent mismatch can be seen in Figure 11f when looking at the practical motivation. Practical generalisation tests are typically aimed at improving models or at being directly informative of a model's applicability. Nonetheless, almost 25% of the practically motivated studies use either artificially partitioned natural data or even fully generated data. To what extent could their conclusions then actually be informative of models applied in practical scenarios? These apparent mismatches between the motivation and the experimental setup exemplify the importance of the motivation axis in our taxonomy – being aware and explicit about it should ensure that the conclusions of a study are indeed informative of the question it claims to answer.

Another interesting observation that can be made from the interactions between motivation and shift locus is that the vast majority of cognitively motivated studies are conducted in a train–test setup. While there are many good reasons for this, conclusions about human generalisation are drawn from a much more varied range of 'experimental setups'. For instance, any experiments done with adults are more similar to finetune train–test or pretrain–test locus than to the train–test locus, as adults have a lifelong experience over which the experimenter has little control beyond participant selection. On the one hand, this suggests that generalisation with a cognitive motivation should perhaps be evaluated more with those loci. On the other hand, it begs the question: for the – previously reported challenging – evaluation of generalisation of LLMs trained on uncontrolled corpora in a pretrain–test setting, could we perhaps take inspiration from how generalisation is evaluated in 'pretrained' humans? While there are, of course, substantial differences between the assumptions that can reasonably be made about the history of a human and the pretraining of an LLM[22], we still believe that input from domain experts that

[21] The observant reader might note that there are, in fact, also several covariate and full shifts with a pretrain-train locus, as well as covariate shifts with a pretrain–test locus. These typically do not represent experiments with LLMS but instead, for instance, consider a multi-stage process for domain adaptation, which also includes a zero-shot comparison.

[22] On the one hand, for a human, some assumptions can be safely made or even verified with a participant – for instance, unless a person has previously participated in a psycholinguistic experiment, we can almost be certain that they have never

have extensively considered human generalisation might be very beneficial to improve generalisation testing in these more challenging setups.

8 Conclusion

While the ability to generalise well – i.e. to successfully transfer skills learned from past experience to new experiences – is considered to be one of the primary desiderata for NLP models, there is very little agreement on what kind of generalisation behaviour modern-age NLP models should exhibit, and under what conditions that should be evaluated. For decades, generalisation has been simply evaluated with random train–test splits. The recent past, however, has seen a number of studies illustrating that models that exhibit near-perfect performances on such i.i.d. splits can sometimes drastically fail in a wide range of scenarios that require different forms of generalisation. This body of work demonstrates the need for more comprehensive generalisation testing, but does not provide much guidance on what that should look like: different papers use different experimental setups, different types of data and entertain even different ideas about what it means for an NLP model to generalise well. As a consequence, even though its importance is almost undisputed, extensive, state-of-the-art generalisation testing is not currently the standard in NLP. With this paper, we aimed to set the first steps towards making it the new status quo.

8.1 Our generalisation taxonomy

We presented a new framework to systematise and understand generalisation research, with the ultimate goal to lay the groundwork for making generalisation testing the new status quo in NLP. The first part of this framework consists of a generalisation taxonomy that can be used to characterise generalisation studies along various dimensions. This taxonomy, which is designed based on an extensive review of generalisation papers in NLP, can be used to critically analyse existing generalisation research and to structure new studies. It contains five nominal axes, that describe *why* the study was executed (the main **motivation** of the study), *what* the study intends to evaluate (the **type** of generalisation they aim to solve), and *how* it does so (the type of **data shift** they are considering, the **source** by which this data shift was obtained, and the **locus** in which the shift is investigated). An overview of our taxonomy is provided in Figure 1; the axes are discussed in §2-6.

8.2 Our analysis

To illustrate the use and usefulness of our taxonomy, we analysed by means of it 449 papers that have the (sub)words generali(s/z)ation or generali(s/z)e in their title or abstract. We hope that researchers will use our taxonomy to design future generalisation studies and to critically and explicitly characterise their experiments. To this end, on our website, we provide an annotation diagram that can be used to design and conceptualise generalisation studies. Through our extensive analysis, we demonstrated that the taxonomy is applicable to a wide range of generalisation studies, and we were able to provide a comprehensive map of the field, observing overall patterns and making suggestions for areas that should be prioritised in the future. In §7, we described the results of this review: we discussed overall patterns on individual axes, as well as interactions between different axes and trends over time – all illustrated with compelling data visualisations. Our most important conclusions and recommendations are:

- The goal of a study is not always perfectly lined up with its experimental design. We advise that future work is explicit about their motivations – which strongly impact what sort of generalisation

conjugated *nonce words*. For an LLM, this is less trivially true, as reports about such human experiments may have been present in their (pre)training data. On the other hand, for an LLM it is possible to inspect the data that they have seen during pretraining, which is evidently not the case for humans.

is even desirable – and should incorporate deliberate assessments to ensure that the experimental setup is aligned with the goal of the study.

- Cross-lingual studies and generalisation studies motivated by fairness goals are underrepresented. We suggest that these areas be given more attention in future work.
- Papers that target similar generalisation questions vary widely in the type of evaluation setup they use. In our view, the field would benefit from more *meta-studies* that consider how the results of experiments with different experimental paradigms compare to each other.
- The vast majority of generalisation studies focuses on only one stage of the modelling pipeline. More work is needed that considers generalisation in all stages of training, to prioritise models whose generalising behaviour persists throughout their training curriculum.
- Recent popular NLP models that can be tested directly for their generalisation from pretraining to testing (e.g. in prompting setups, without any further model training) have often been evaluated without considering the relationship between the (pre)training and the test data. We envisage that this is due to the fact that generalisation is particularly difficult to assess when large uncontrolled training data is involved, and we suggest that inspiration might be taken from how generalisation is evaluated in experiments with adult humans, where control and access to the "pretraining" data of a participant are unattainable.

Along with this paper, we also launch a website with a set of visualisation tools and the possibility to browse through our review to find studies with specific features, as well as relevant paper references. While the review and conclusions presented in this paper are necessarily static, we commit to keeping the entries on the website up to date when new papers on generalisation are published and we encourage researchers to engage with our online dynamic review by submitting both new studies and existing studies we might have missed – through the contributions page of our website.

8.3 Future work

By providing a systematic framework and set of concrete (online) tools to allow for a structured understanding of generalisation, we believe we have set the necessary first step towards making state-of-the-art generalisation testing the new status quo in NLP. Our work is thus by no means the end of the road. While our taxonomy can make future generalisation research in NLP more *comparable*, *structured* and *carefully designed*, and while our survey suggests promising research directions, this work does not provide standardised data or procedures for generalisation testing. We envision that important generalisation tests should be hosted on a shared platform, along with a leaderboard to make generalisation testing more accessible and transparent. A large community of NLP researchers and domain experts should determine which tests to prioritise. Lastly, in the same way that our thoughts on how generalisation should be evaluated have evolved with our models in the past, it will likely continue to do so in the future. What we consider important to evaluate now might change next year, and when models get better at setups considered difficult now, we might discover new types of generalisation that we had not thought of before. How we evaluate models should be reflective of that, and which tests are prioritised should thus evolve along with our models and knowledge. Ideally, all of those aspects should be incorporated in the next steps towards making state-of-the-art generalisation testing the new status quo for any new model that is proposed, and we look forward to working on it.

9 Limitations

Designing a coherent, consistent, and at the same time, usable taxonomy of generalisation research in NLP is a non-trivial task, which required substantial discussion among the authors. In this section, we report the main decisional trade-offs of our work, concerning the definition of the taxonomy, the annotation process and the selection of papers to review.

9.1 Taxonomy design: the axes and their values

We designed this taxonomy by ensuring that the selected set of axes and axis values would highlight theoretically important but also practically functional distinctions between generalisation studies – yet our selection comes with limitations. One such limitation is that the axis values are relatively coarse. This avoids fragmentation in the analysis and allows to draw higher-level conclusions, but sometimes also groups together papers that could be regarded separately. An already discussed example are the studies with a pretrain-train locus, which by definition all share that they include more than one training stage and investigate generalisation in the first one. This category thus contains both papers that use a general-purpose pretraining objective and then finetune on different tasks and studies whose finetuning objective matches the pretraining objectives (e.g. studies that consider domain-adaptation in a continual learning setup). While those differences are – at least in part – reflected on other axes, in some cases it might be helpful to distinguish those two cases more explicitly.

Something similar occurs on the shift type axis. Firstly, when there are multiple shifts, we do not currently distinguish between all possible combinations of individual shift types. Given the relatively low number of studies that actually consider multiple shifts, we prioritised intelligibility over completeness, but if the number of multiple-shift studies increases in the future, it could become useful to indicate all individual shift types in the case of studies with multiple shifts. Secondly, while the three formal shift types that we consider are statistically well-grounded, shifts of the same type can still largely vary. Whether the distance between two distributions is small or large might make a substantial difference for the difficulty of the generalisation problem, which is something that is currently not reflected in our taxonomy. Although quantifying differences between distributions is often problematic in practice, we believe that adjusting the taxonomy to capture the difficulty of generalising to a particular shift can be helpful in the future. More generally, we imagine that future experimental paradigms might call for the addition of values on some of the axes, or even the addition of new axes.

9.2 Annotation: axes values in practice

In the description of the axes and their different values, we aimed to be as comprehensive and precise as possible. In practice, however, there are always cases in which the actual category of a paper is debatable. Sometimes this occurs because the paper itself is not clear about what exactly it attempts to evaluate or about its motivation; we hope that our taxonomy will reduce the number of such cases in the future. In other instances, it is simply difficult to apply some concepts or distinctions, in spite of their theoretical sharpness, to concrete studies. A clear example of this challenge is the shift type. In theory, $p(x)$, $p(y|x)$ and $p(y)$ are clearly defined concepts; in practice, it is usually impossible to estimate the actual difference between two (natural) distributions. Some researchers might even argue that, in practice, train and test sets are virtually always distributionally different. For the purpose of systematising generalisation testing and characterising experiments, however, this is not a useful observation. In our taxonomy design and annotations, we aimed to make distinctions that we deemed useful, rather than relying on "true" but unknown differences between distributions.

9.3 Paper selection

To ensure that our selection of papers was not biased toward works already known by the authors, we automatically selected a large number of papers from the ACL anthology by searching for generalisation keywords in the abstract and title. While this resulted in a relatively large amount of papers, there are likely papers about generalisation that we did not retrieve with this approach. As mentioned earlier

(§7.2.1), we suspect that papers about cross-lingual generalisation and papers with a fairness motivation may require a different set of keywords. We hope that researchers will take the effort to inform us about generalisation papers that we may have missed, to guarantee that the selection of surveyed papers is as complete as possible.

Aside from unintentionally missed papers, we also deliberately excluded a few types of papers. We did not include any studies that considered more than one modality. While we believe they are interesting to consider from a generalisation perspective, they are also more difficult to characterise within a single taxonomy, as they involve more distributions (with sometimes very different support) and thus more distribution shifts. We consider including such papers a compelling step for future work. Another set of papers that we excluded are those that do not conduct behavioural experiments but look at the generalisability of representations (e.g. probing papers). We do not see any a priori reason that they could not be characterised with our taxonomy, and we believe this would be a valuable enterprise. In particular, although marking the difference between behavioural and representational experiments might require updating the taxonomy, a comparison of behavioural and representational experiments with the same axis values might make for an interesting meta-study.

9.4 Is generalisation always necessary?

A last critical observation that we would like to make is that our work builds on the assumption that strong generalisation skills are considered crucial for models of NLP. While we generally believe this to be true, there might be cases where generalisation is not in fact needed. Provocatively, one could argue that for LLMs trained on extremely large English data sets, practically speaking the vast majority of scenarios that one might want to use the model for is actually close to i.i.d. and that more complex forms of generalisation are thus not needed. We abstain from judging whether and when this holds, but argue that if a researcher believes that their setup requires no generalisation, they should clearly state so and explain why they believe that to be the case.

Acknowledgements

We thank Adina Williams, Armand Joulin, Elia Bruni, Lucas Weber, Robert Kirk and Sebastian Riedel for providing us feedback on various stages of this draft, and Gary Marcus for providing detailed feedback on the final draft of this paper. We thank Elte Hupkes for making the app that allows searching through references, and we thank Daniel Haziza, Ece Takmaz and Maria Ryskina for other contributions to the website.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and

Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Prabal Agarwal, Jannik Strötgen, Luciano del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. diaNED: Time-aware named entity disambiguation for diachronic corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693, Melbourne, Australia. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Maruan Al-Shedivat and Ankur Parikh. 2019. Consistency by agreement in zero-shot neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1184–1197, Minneapolis, Minnesota. Association for Computational Linguistics.

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2022. Naturalistic causal probing for morpho-syntax. *arXiv preprint arXiv:2205.07043*.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2022. Ext5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations*.

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giri Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Mona T. Diab, Zornitsa Kozareva, and Ves Stoyanov. 2021. Efficient large scale language modeling with mixtures of experts. *CoRR*, abs/2112.10684.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *CoRR*, abs/2106.08694.

Hanna Behnke, Marina Fomicheva, and Lucia Specia. 2022. Bias mitigation in machine translation quality estimation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1475–1487, Dublin, Ireland. Association for Computational Linguistics.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Jean Berko. 1958. The child's learning of English morphology. *Word*, 14(2-3):150–177.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

Prajjwal Bhargava, Aleksandr Drozd, and Anna Rogers. 2021. Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pages 125–135, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. 2020. Continual lifelong learning in natural language processing: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6523–6541, Barcelona, Spain (Online). International Committee on Computational Linguistics.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2017. A dataset and classifier for recognizing social media English. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 56–61, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019. *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Association for Computational Linguistics, Florence, Italy.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R Bowman, Christopher D Manning, and Christopher Potts. 2015b. Tree-structured composition in neural networks without tree-structured architectures. In *Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches*, pages 37–42.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Basilio Calderone, Nabil Hathout, and Olivier Bonami. 2021. Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 274–282, Online. Association for Computational Linguistics.

Rahma Chaabouni, Roberto Dessì, and Eugene Kharitonov. 2021. Can transformers jump around right in natural language? assessing performance transfer from SCAN. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 136–148, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Michael John Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191, Santa Cruz, California, USA. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 160–167. ACM.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. Are we there yet? encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877, Florence, Italy. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. 2022. Generalized quantifiers as a source of error in multilingual NLU benchmarks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4875–4893, Seattle, United States. Association for Computational Linguistics.

Marco Damonte and Emilio Monti. 2021. One semantic parser to parse them all: Sequence to sequence multi-task learning on semantic parsing datasets. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 173–184, Online. Association for Computational Linguistics.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Verna Dankers, Anna Langedijk, Kate McCurdy, Adina Williams, and Dieuwke Hupkes. 2021. Generalising to German plural noun classes, from the perspective of a recurrent neural network. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 94–108, Online. Association for Computational Linguistics.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Erenay Dayanik and Sebastian Padó. 2021. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61, Online. Association for Computational Linguistics.

Andrea De Varda and Roberto Zamparelli. 2022. Multilingualism encourages recursion: a transfer study with mBERT. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 1–10, Seattle, Washington. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Yann Dubois, Gautier Dagan, Dieuwke Hupkes, and Elia Bruni. 2020. Location Attention for Extrapolation to Longer Sequences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 403–413, Online. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring causal effects of data statistics on language model's 'factual' predictions. *CoRR*, abs/2207.14251.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021a. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021b. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 58–63, Valencia, Spain. Association for Computational Linguistics.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-SQL evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360, Melbourne, Australia. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation.

Daniel Fried, Nikita Kitaev, and Dan Klein. 2019. Cross-domain generalization of neural constituency parsers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating NER for Twitter #drift. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2544–2547, Reykjavik, Iceland. European Language Resources Association (ELRA).

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2786–2791, Florence, Italy. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Xiaochuang Han and Yulia Tsvetkov. 2022. ORCA: interpreting prompted language models via locating supporting data evidence in the ocean of pretraining data. *CoRR*, abs/2205.12600.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Yu He, Jianxin Li, Yangqiu Song, Mutian He, and Hao Peng. 2018. Time-evolving text classification with deep neural networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2241–2247. International Joint Conferences on Artificial Intelligence Organization.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.

Dieuwke Hupkes, Sara , and Willem Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Dieuwke Hupkes. 2020. Hierarchy and interpretability in neural models of language processing.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2020. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intellgence Research*, 67:757–795.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Shafiq Joty, Hassan Sajjad, Nadir Durrani, Kamla Al-Mannai, Ahmed Abdelali, and Stephan Vogel. 2015. How to avoid unwanted pregnancies: Domain adaptation using neural network models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1259–1270, Lisbon, Portugal. Association for Computational Linguistics.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Pride Kavumba, Ryo Takahashi, and Yusuke Oda. 2022. Are prompt-based models clueless? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2333–2352, Dublin, Ireland. Association for Computational Linguistics.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Temuulen Khishigsuren, Gábor Bella, Khuyagbaatar Batsuren, Abed Alhakim Freihat, Nandu Chandran Nair, Amarsanaa Ganbold, Hadi Khalilia, Yamini Chandrashekar, and Fausto Giunchiglia. 2022. Using linguistic typology to enrich multilingual lexicons: the case of lexical gaps in kinship. *CoRR*, abs/2204.05049.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. 2021. A survey of generalisation in deep reinforcement learning. *CoRR*, abs/2111.09794.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic. Association for Computational Linguistics.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4487–4499.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

Yair Lakretz, Theo Desbordes, Dieuwke Hupkes, and Stanislas Dehaene. 2021a. Causal transformers perform below chance on recursive nested constructions, unlike humans. *CoRR*, abs/2110.07240.

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021b. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, 213:104699. Special Issue in Honour of Jacques Mehler, Cognition's founding editor.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Ian Lane, Tatsuya Kawahara, Tomoko Matsui, and Satoshi Nakamura. 2007. Out-of-domain utterance detection using classification confidences of multiple topics. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):150–161.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, pages 7315–7330, Online. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Belinda Li, Jane Yu, Madian Khabsa, Luke Zettlemoyer, Alon Halevy, and Jacob Andreas. 2022. Quantifying adaptability in pre-trained language models with 500 tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4696–4715, Seattle, United States. Association for Computational Linguistics.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021a. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Jane S.Y. Li and Colin Wilson. 2021. Leveraging paradigmatic information in inflection acceptability prediction: The JHU-SFU submission to SIGMORPHON shared task 0.2. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 289–294, Online. Association for Computational Linguistics.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021b. On compositional generalization of neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4767–4780, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Jindřich Libovický, Helmut Schmid, and Alexander Fraser. 2022. Why don't people use character-level machine translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2470–2485, Dublin, Ireland. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019a. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2022. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749, Dublin, Ireland. Association for Computational Linguistics.

Linqing Liu, Patrick S. H. Lewis, Sebastian Riedel, and Pontus Stenetorp. 2021. Challenges in generalization in open domain question answering. *CoRR*, abs/2109.01156.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Florian Ludwig, Klara Dolos, Torsten Zesch, and Eleanor Hobley. 2022. Improving generalization of hate speech detection systems to novel target groups via domain adaptation. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 29–39, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jan Lukes and Anders Søgaard. 2018. Sentiment analysis under temporal shift. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 65–71, Brussels, Belgium. Association for Computational Linguistics.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 276–283, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Andrey Malinin, Neil Band, Yarin Gal, Mark J. F. Gales, Alexander Ganshin, German Chesnokov, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panagiotis Tigas, and Boris Yangel. 2021. Shifts: A dataset of real distributional shift across multiple large-scale tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Robert Malouf. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

Gary Marcus. 2018. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631.

Gary F. Marcus. 1998. Rethinking eliminative connectionism. *Cognitive Psychology*, 37(3):243–282.

Gary F Marcus. 1999. Connectionism: with or without rules?: Response to jl mcclelland and dc plaut (1999). *Trends in Cognitive Sciences*, 3(5):168–170.

Gary F Marcus. 2003. *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Gary F Marcus, Ursula Brinkmann, Harald Clahsen, Richard Wiese, and Steven Pinker. 1995. German inflection: The exception that proves the rule. *Cognitive psychology*, 29(3):189–256.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

James L McClelland and David C Plaut. 1999. Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences*, 3(5):166–168.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756, Online. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. 2012. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Quan Hung Tran, Trung Bui, Walter Chang, and Ndapa Nakashole. 2020. Rethinking self-attention: Towards interpretability in neural parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 731–742, Online. Association for Computational Linguistics.

Mathijs Mul and Willem Zuidema. 2019. Siamese recurrent networks learn first-order logic reasoning and exhibit zero-shot compositional generalization. In *CoRR, abs/1906.00180*.

Benjamin Muller, Luca Soldaini, Rik Koncel-Kedziorski, Eric Lind, and Alessandro Moschitti. 2021. Cross-lingual genqa: Open-domain question answering with answer sentence generation.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 90–97, Hong Kong, China. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021a. Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021b. Sigmorphon 2021 shared task on morphological reinflection part 2: Are we there yet? A shared task on cognitively plausible morphological inflection.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *arXiv preprint arXiv:1608.07836*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Vikas Raunak, Siddharth Dalmia, Vivek Gupta, and Florian Metze. 2020. On long-tailed phenomena in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3088–3095, Online. Association for Computational Linguistics.

Vikas Raunak, Vaibhav Kumar, Florian Metze, and Jaimie Callan. 2019. On compositionality in neural machine translation. In *NeurIPS 2019 Context and Compositionality in Biological and Artificial Neural Systems Workshop*.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning. *CoRR*, abs/2202.07206.

Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7605–7617, Online. Association for Computational Linguistics.

Roni Rosenfeld. 1996. A maximum entropy approach to adaptive statistical language modeling.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seonghan Ryu, Seokhwan Kim, Junhwi Choi, Hwanjo Yu, and Gary Geunbae Lee. 2017. Neural sentence embedding using only in-domain sentences for out-of-domain sentence detection in dialog systems. *Pattern Recognition Letters*, 88:26–32.

Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. 2018. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 714–718, Brussels, Belgium. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Federico Sangati and Willem Zuidema. 2011. Accurate parsing with compact tree-substitution grammars: Double-DOP. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 84–95, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Jürgen Schmidhuber. 1990. Towards compositional learning in dynamic networks.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Online. Association for Computational Linguistics.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Amos Storkey. 2009. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30:3–28.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. 2020. LAMOL: LAnguage MOdeling for Lifelong Language Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Gábor Szolnok, Botond Barta, Dorina Lakatos, and Judit Ács. 2021. Bme submission for sigmorphon 2021 shared task 0. a three step training approach with data augmentation for morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 268–273, Online. Association for Computational Linguistics.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online harassment*, pages 29–55. Springer.

Aarne Talman and Stergios Chatzikyriakidis. 2019. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3566–3572, Hong Kong, China. Association for Computational Linguistics.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sara Veldhoen, Dieuwke Hupkes, and Willem Zuidema. 2016. Diagnostic classifiers: Revealing how neural networks process hierarchical structure. In *Proceedings of the NIPS2016 Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Bailin Wang, Mirella Lapata, and Ivan Titov. 2021. Meta-learning for domain generalization in semantic parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–379, Online. Association for Computational Linguistics.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017a. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 560–566, Vancouver, Canada. Association for Computational Linguistics.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017b. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Benchmarking generalization via in-context instructions on 1,600+ language tasks.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. 2021. Language modelling as a multi-task problem. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2049–2060, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

(Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Wilson and Jane S.Y. Li. 2021. Were we there already? applying minimal generalization to the sigmorphon-unimorph shared task on cognitively plausible morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 283–291, Online. Association for Computational Linguistics.

Francis CK Wong and William SY Wang. 2007. Generalisation towards combinatorial productivity in language acquisition by simple recurrent networks. In *2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems*, pages 139–144. IEEE.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *CoRR*, abs/2201.05966.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, and Kentaro Inui. 2021. Exploring transitivity in neural NLI models through veridicality. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 920–934, Online. Association for Computational Linguistics.

Kaiyu Yang and Jia Deng. 2020. Strongly incremental constituency parsing with graph neural networks. *Advances in Neural Information Processing Systems*, 33:21687–21698.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Zhou, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. 2021. Hidden biases in unreliable news detection datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2482–2492, Online. Association for Computational Linguistics.

Zhong Zhou, Matthias Sperber, and Alexander Waibel. 2018. Massively parallel cross-lingual learning in low-resource target language translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 232–243, Brussels, Belgium. Association for Computational Linguistics.

List of publications included in our review

Adelani David, Ruiter Dana, Alabi Jesujoba, Adebonojo Damilola, Ayeni Adesina, Adeyemi Mofe, Awokoya Ayodele Esther, España-Bonet Cristina. The Effect of Domain and Diacritics in Yoruba–English Neural Machine Translation // Proceedings of Machine Translation Summit XVIII: Research Track. Virtual: Association for Machine Translation in the Americas, VIII 2021. 61–75.

Aghajanyan Armen, Gupta Anchit, Shrivastava Akshat, Chen Xilun, Zettlemoyer Luke, Gupta Sonal. Muppet: Massive Multi-task Representations with Pre-Finetuning // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 5799–5811.

Ahamad Afroz, Anand Ankit, Bhargava Pranesh. AccentDB: A Database of Non-Native English Accents to Assist Neural Speech Recognition // Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, V 2020. 5351–5358.

Ahmad Wasi, Li Haoran, Chang Kai-Wei, Mehdad Yashar. Syntax-augmented Multilingual BERT for Cross-lingual Transfer // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 4538–4554.

Ahuja Ojas, Xu Jiacheng, Gupta Akshay, Horecka Kevin, Durrett Greg. ASPECTNEWS: Aspect-Oriented Summarization of News Documents // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 6494–6506.

Ajjour Yamen, Chen Wei-Fan, Kiesel Johannes, Wachsmuth Henning, Stein Benno. Unit Segmentation of Argumentative Texts // Proceedings of the 4th Workshop on Argument Mining. Copenhagen, Denmark: Association for Computational Linguistics, IX 2017. 118–128.

Akyurek Ekin, Andreas Jacob. Lexicon Learning for Few Shot Sequence Modeling // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 4934–4946.

Al-Shedivat Maruan, Parikh Ankur. Consistency by Agreement in Zero-Shot Neural Machine Translation // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 1184–1197.

Alkiek Kenan, Zhang Bohan, Jurgens David. Classification without (Proper) Representation: Political Heterogeneity in Social Media and Its Implications for Classification and Behavioral Analysis // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 504–522.

Allaway Emily, McKeown Kathleen. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 8913–8931.

Allaway Emily, Srikanth Malavika, McKeown Kathleen. Adversarial Learning for Zero-Shot Stance Detection on Social Media // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 4756–4767.

Amin Saadullah, Pokaratsiri Goldstein Noon, Wixted Morgan, Garcia-Rudolph Alejandro, Martínez-Costa Catalina, Neumann Guenter. Few-Shot Cross-lingual Transfer for Coarse-grained De-identification of Code-Mixed Clinical Texts // Proceedings of the 21st Workshop on Biomedical Language Processing. Dublin, Ireland: Association for Computational Linguistics, V 2022. 200–211.

Ammanabrolu Prithviraj, Jia Renee, Riedl Mark. Situated Dialogue Learning through Procedural Environment Generation // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 8099–8116.

Ammar Waleed, Mulcaire George, Ballesteros Miguel, Dyer Chris, Smith Noah A. Many Languages, One Parser // Transactions of the Association for Computational Linguistics. 2016. 4. 431–444.

Andrews Nicholas, Bishop Marcus. Learning Invariant Representations of Social Media Users // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 1684–1695.

Angelidis Stefanos, Frermann Lea, Marcheggiani Diego, Blanco Roi, Màrquez Lluís. Book QA: Stories of Challenges and Opportunities // Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Hong Kong, China: Association for Computational Linguistics, XI 2019. 78–85.

Aribandi Vamsi, Tay Yi, Schuster Tal, Rao Jinfeng, Zheng Huaixiu Steven, Mehta Sanket Vaibhav, Zhuang Honglei, Tran Vinh Q., Bahri Dara, Ni Jianmo, Gupta Jai, Hui Kai, Ruder Sebastian, Metzler Donald.

ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning // International Conference on Learning Representations. 2022.

Artetxe Mikel, Bhosale Shruti, Goyal Naman, Mihaylov Todor, Ott Myle, Shleifer Sam, Lin Xi Victoria, Du Jingfei, Iyer Srinivasan, Pasunuru Ramakanth, Anantharaman Giri, Li Xian, Chen Shuohui, Akin Halil, Baines Mandeep, Martin Louis, Zhou Xing, Koura Punit Singh, O'Horo Brian, Wang Jeff, Zettlemoyer Luke, Diab Mona T., Kozareva Zornitsa, Stoyanov Ves. Efficient Large Scale Language Modeling with Mixtures of Experts // CoRR. 2021. abs/2112.10684.

Artetxe Mikel, Ruder Sebastian, Yogatama Dani. On the Cross-lingual Transferability of Monolingual Representations // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 4623–4637.

Atwell Katherine, Sicilia Anthony, Hwang Seong Jae, Alikhani Malihe. The Change that Matters in Discourse Parsing: Estimating the Impact of Domain Shift on Parser Error // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 824–845.

Auersperger Michal, Pecina Pavel. Solving SCAN Tasks with Data Augmentation and Input Embeddings // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). Held Online: INCOMA Ltd., IX 2021. 86–91.

Bahri Dara, Mobahi Hossein, Tay Yi. Sharpness-Aware Minimization Improves Language Model Generalization // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 7360–7371.

Bai He, Wang Tong, Sordoni Alessandro, Shi Peng. Better Language Model with Hypernym Class Prediction // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 1352–1362.

Bandyopadhyay Saptarashmi, Zhao Tianyang. Natural Language Response Generation from SQL with Generalization and Back-translation // Proceedings of the First Workshop on Interactive and Executable Semantic Parsing. Online: Association for Computational Linguistics, XI 2020. 46–49.

Banerjee Pratyay, Baral Chitta. Self-Supervised Knowledge Triplet Learning for Zero-Shot Question Answering // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 151–162.

Bansal Trapit, Jha Rishikesh, McCallum Andrew. Learning to Few-Shot Learn Across Diverse Natural Language Classification Tasks // Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, XII 2020a. 5108–5123.

Bansal Trapit, Jha Rishikesh, Munkhdalai Tsendsuren, McCallum Andrew. Self-Supervised Meta-Learning for Few-Shot Natural Language Classification Tasks // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020b. 522–534.

Bari M Saiful, Haider Batool, Mansour Saab. Nearest Neighbour Few-Shot Learning for Cross-lingual Classification // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 1745–1753.

Bartolo Max, Roberts Alastair, Welbl Johannes, Riedel Sebastian, Stenetorp Pontus. Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension // Transactions of the Association for Computational Linguistics. 2020. 8. 662–678.

Bartolo Max, Thrush Tristan, Jia Robin, Riedel Sebastian, Stenetorp Pontus, Kiela Douwe. Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 8830–8848.

Bastings Jasmijn, Baroni Marco, Weston Jason, Cho Kyunghyun, Kiela Douwe. Jump to better conclusions: SCAN both left and right // Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium: Association for Computational Linguistics, XI 2018. 47–55.

Basu Roy Chowdhury Somnath, Brahman Faeze, Chaturvedi Snigdha. Is Everything in Order? A Simple Way to Order Sentences // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 10769–10779.

Basu Roy Chowdhury Somnath, M Annervaz, Dukkipati Ambedkar. Instance-based Inductive Deep Transfer Learning by Cross-Dataset Querying with Locality Sensitive Hashing // Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019). Hong Kong, China: Association for Computational Linguistics, XI 2019. 183–191.

Bau D. Anthony, Andreas Jacob. How Do Neural Sequence Models Generalize? Local and Global Cues for Out-of-Distribution Prediction // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 5513–5526.

Baumann Antonia. Multilingual Language Models for Named Entity Recognition in German and English // Proceedings of the Student Research Workshop Associated with RANLP 2019. Varna, Bulgaria: INCOMA Ltd., IX 2019. 21–27.

Behnke Hanna, Fomicheva Marina, Specia Lucia. Bias Mitigation in Machine Translation Quality Estimation // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 1475–1487.

Betz Gregor, Voigt Christian, Richardson Kyle. Critical Thinking for Language Models // Proceedings of the 14th International Conference on Computational Semantics (IWCS). Groningen, The Netherlands (online): Association for Computational Linguistics, VI 2021. 63–75.

Bhargava Prajjwal, Drozd Aleksandr, Rogers Anna. Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics // Proceedings of the Second Workshop on Insights from Negative Results in NLP. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 125–135.

Bingel Joachim, Bjerva Johannes. Cross-lingual complex word identification with multitask learning // Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. New Orleans, Louisiana: Association for Computational Linguistics, VI 2018. 166–174.

Bjerva Johannes, Kementchedjhieva Yova, Cotterell Ryan, Augenstein Isabelle. A Probabilistic Generative Model of Linguistic Typology // Proceedings of the 2019 Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 1529–1540.

Bodapati Sravan, Yun Hyokun, Al-Onaizan Yaser. Robustness to Capitalization Errors in Named Entity Recognition // Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). Hong Kong, China: Association for Computational Linguistics, XI 2019. 237–242.

Borkan Daniel, Dixon Lucas, Sorensen Jeffrey, Thain Nithum, Vasserman Lucy. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification // Companion Proceedings of The 2019 World Wide Web Conference. New York, NY, USA: Association for Computing Machinery, 2019. 491–500. (WWW '19).

Bowman Samuel R, Manning Christopher D, Potts Christopher. Tree-structured composition in neural networks without tree-structured architectures // Proceedings of the 2015th International Conference on Cognitive Computation: Integrating Neural and Symbolic Approaches. 2015. 37–42.

Bragg Jonathan, Cohan Arman, Lo Kyle, Beltagy Iz. FLEX: Unifying Evaluation for Few-Shot NLP // Advances in Neural Information Processing Systems. 34. 2021. 15787–15800.

Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared D, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, others . Language models are few-shot learners // Advances in neural information processing systems. 2020. 33. 1877–1901.

Büyüköz Berfu, Hürriyetoğlu Ali, Özgür Arzucan. Analyzing ELMo and DistilBERT on Socio-political News Classification // Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020. Marseille, France: European Language Resources Association (ELRA), V 2020. 9–18.

Cabrera-Diego Luis Adrián, Moreno Jose G., Doucet Antoine. Using a Frustratingly Easy Domain and Tagset Adaptation for Creating Slavic Named Entity Recognition Systems // Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. Kiyv, Ukraine: Association for Computational Linguistics, IV 2021. 98–104.

Castro Ferreira Thiago, Lee Chris van der, Miltenburg Emiel van, Krahmer Emiel. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 552–562.

Cattan Oralie, Rosset Sophie, Servan Christophe. On the cross-lingual transferability of multilingual prototypical models across NLU tasks // Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing. Online: Association for Computational Linguistics, VIII 2021. 36–43.

Cengiz Cemil, Yuret Deniz. Joint Training with Semantic Role Labeling for Better Generalization in Natural Language Inference // Proceedings of the 5th Workshop on Representation Learning for NLP. Online: Association for Computational Linguistics, VII 2020. 78–88.

Cerda-Mardini Patricio, Araujo Vladimir, Soto Álvaro. Translating Natural Language Instructions for Behavioral Robot Navigation with a Multi-Head Attention Mechanism // Proceedings of the The Fourth Widening Natural Language Processing Workshop. Seattle, USA: Association for Computational Linguistics, VII 2020. 96–98.

Chaabouni Rahma, Dessì Roberto, Kharitonov Eugene. Can Transformers Jump Around Right in Natural Language? Assessing Performance Transfer from SCAN // Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 136–148.

Chalkidis Ilias, Jana Abhik, Hartung Dirk, Bommarito Michael, Androutsopoulos Ion, Katz Daniel, Aletras Nikolaos. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 4310–4330.

Chaudhary Aditi, Zhou Chunting, Levin Lori, Neubig Graham, Mortensen David R., Carbonell Jaime. Adapting Word Embeddings to New Languages with Morphological and Phonological Subword Representations // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, X-XI 2018. 3285–3295.

Chauhan Kumud. NEU at WNUT-2020 Task 2: Data Augmentation To Tell BERT That Death Is Not Necessarily Informative // Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). Online: Association for Computational Linguistics, XI 2020. 440–443.

Chen Jiaao, Shen Dinghan, Chen Weizhu, Yang Diyi. HiddenCut: Simple Data Augmentation for Natural Language Understanding with Better Generalizability // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021a. 4380–4390.

Chen Jiaao, Yang Diyi. Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 1380–1391.

Chen Jiawei, Lin Hongyu, Han Xianpei, Sun Le. Honey or Poison? Solving the Trigger Curse in Few-shot Event Detection via Causal Intervention // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021b. 8078–8088.

Chen Jifan, Durrett Greg. Robust Question Answering Through Sub-part Alignment // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 1251–1263.

Chen Qian, Zhu Xiaodan, Ling Zhen-Hua, Wei Si, Jiang Hui, Inkpen Diana. Recurrent Neural Network-Based Sentence Encoder with Gated Attention for Natural Language Inference // Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP. Copenhagen, Denmark: Association for Computational Linguistics, IX 2017. 36–40.

Chen Wenhu, Su Yu, Yan Xifeng, Wang William Yang. KGPT: Knowledge-Grounded Pre-Training for Data-to-Text Generation // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020a. 8635–8648.

Chen Xilun, Ghoshal Asish, Mehdad Yashar, Zettlemoyer Luke, Gupta Sonal. Low-Resource Domain Adaptation for Compositional Task-Oriented Semantic Parsing // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020b. 5090–5100.

Chen Yen-Chun, Bansal Mohit. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, VII 2018. 675–686.

Chen Yiran, Liu Pengfei, Zhong Ming, Dou Zi-Yi, Wang Danqing, Qiu Xipeng, Huang Xuanjing. CDEvalSumm: An Empirical Study of Cross-Dataset Evaluation for Neural Summarization Systems // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020c. 3679–3691.

Chen Yunmo, Chen Tongfei, Ebner Seth, White Aaron Steven, Van Durme Benjamin. Reading the Manual: Event Extraction as Definition Comprehension // Proceedings of the Fourth Workshop on Structured Prediction for NLP. Online: Association for Computational Linguistics, XI 2020d. 74–83.

Chen Zhi, Chen Lu, Zhao Yanbin, Cao Ruisheng, Xu Zihan, Zhu Su, Yu Kai. ShadowGNN: Graph Projection Neural Network for Text-to-SQL Parser // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021c. 5567–5577.

Chen Zhiyu, Eavani Harini, Chen Wenhu, Liu Yinyin, Wang William Yang. Few-Shot NLG with Pre-Trained Language Model // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020e. 183–190.

Cheng Hao, Liu Xiaodong, Pereira Lis, Yu Yaoliang, Gao Jianfeng. Posterior Differential Regularization with f-divergence for Improving Model Robustness // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 1078–1089.

Cheng Yong, Bapna Ankur, Firat Orhan, Cao Yuan, Wang Pidong, Macherey Wolfgang. Multilingual Mix: Example Interpolation Improves Multilingual Neural Machine Translation // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 4092–4102.

Chiang David, Cholak Peter. Overcoming a Theoretical Limitation of Self-Attention // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 7654–7664.

Choudhary Chinmay. Improving the Performance of UDify with Linguistic Typology Knowledge // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. Online: Association for Computational Linguistics, VI 2021. 38–60.

Chowdhery Aakanksha, Narang Sharan, Devlin Jacob, Bosma Maarten, Mishra Gaurav, Roberts Adam, Barham Paul, Chung Hyung Won, Sutton Charles, Gehrmann Sebastian, Schuh Parker, Shi Kensen, Tsvyashchenko Sasha, Maynez Joshua, Rao Abhishek, Barnes Parker, Tay Yi, Shazeer Noam, Prabhakaran Vinodkumar, Reif Emily, Du Nan, Hutchinson Ben, Pope Reiner, Bradbury James, Austin Jacob, Isard Michael, Gur-Ari Guy, Yin Pengcheng, Duke Toju, Levskaya Anselm, Ghemawat Sanjay, Dev Sunipa, Michalewski Henryk, Garcia Xavier, Misra Vedant, Robinson Kevin, Fedus Liam, Zhou Denny, Ippolito Daphne, Luan David, Lim Hyeontaek, Zoph Barret, Spiridonov Alexander, Sepassi Ryan, Dohan David, Agrawal Shivani, Omernick Mark, Dai Andrew M., Pillai Thanumalayan Sankaranarayana, Pellat Marie, Lewkowycz Aitor, Moreira Erica, Child Rewon, Polozov Oleksandr, Lee Katherine, Zhou Zongwei, Wang Xuezhi, Saeta Brennan, Diaz Mark, Firat Orhan, Catasta Michele, Wei Jason, Meier-Hellstern Kathy, Eck Douglas, Dean Jeff, Petrov Slav, Fiedel Noah. Palm: Scaling language modeling with pathways // CoRR. 2022. abs/2204.02311.

Chowdhury Shammur Absar, Mubarak Hamdy, Abdelali Ahmed, Jung Soon-gyo, Jansen Bernard J., Salminen Joni. A Multi-Platform Arabic News Comment Dataset for Offensive Language Detection // Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, V 2020. 6203–6212.

Collobert Ronan, Weston Jason. A unified architecture for natural language processing: deep neural networks with multitask learning // Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008. 307. 2008. 160–167. (ACM International Conference Proceeding Series).

Conforti Costanza, Berndt Jakob, Basaldella Marco, Pilehvar Mohammad Taher, Giannitsarou Chryssi, Toxvaerd Flavio, Collier Nigel. Adversarial Training for News Stance Detection: Leveraging Signals from a Multi-Genre Corpus. // Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. Online: Association for Computational Linguistics, IV 2021a. 1–7.

Conforti Costanza, Berndt Jakob, Pilehvar Mohammad Taher, Giannitsarou Chryssi, Toxvaerd Flavio, Collier Nigel. Synthetic Examples Improve Cross-Target Generalization: A Study on Stance Detection on a Twitter corpus. // Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Online: Association for Computational Linguistics, IV 2021b. 181–187.

Conklin Henry, Wang Bailin, Smith Kenny, Titov Ivan. Meta-Learning to Compositionally Generalize // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 3322–3335.

Costa-jussà Marta R., Cross James, Çelebi Onur, Elbayad Maha, Heafield Kenneth, Heffernan Kevin, Kalbassi Elahe, Lam Janice, Licht Daniel, Maillard Jean, Sun Anna, Wang Skyler, Wenzek Guillaume, Youngblood Al, Akula Bapi, Barrault Loïc, Gonzalez Gabriel Mejia, Hansanti Prangthip, Hoffman John, Jarrett Semarley, Sadagopan Kaushik Ram, Rowe Dirk, Spruit Shannon, Tran Chau, Andrews Pierre, Ayan Necip Fazil, Bhosale Shruti, Edunov Sergey, Fan Angela, Gao Cynthia, Goswami Vedanuj, Guzmán Francisco, Koehn Philipp, Mourachko Alexandre, Ropers Christophe, Saleem Safiyyah, Schwenk Holger, Wang Jeff. No Language Left Behind: Scaling Human-Centered Machine Translation // CoRR. 2022. abs/2207.04672.

Cruz Jan Christian Blaise, Tan Julianne Agatha, Cheng Charibeth. Localization of Fake News Detection via Multitask Transfer Learning // Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, V 2020. 2596–2604.

Csordás Róbert, Irie Kazuki, Schmidhuber Juergen. The Devil is in the Detail: Simple Tricks Improve Systematic Generalization of Transformers // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 619–634.

Cui Ruixiang, Hershcovich Daniel, Søgaard Anders. Generalized Quantifiers as a Source of Error in Multilingual NLU Benchmarks // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, VII 2022. 4875–4893.

Czarnowska Paula, Ruder Sebastian, Grave Edouard, Cotterell Ryan, Copestake Ann. Don't Forget the Long Tail! A Comprehensive Analysis of Morphological Generalization in Bilingual Lexicon Induction // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 974–983.

Dalvi Bhavana, Jansen Peter, Tafjord Oyvind, Xie Zhengnan, Smith Hannah, Pipatanangkura Leighanna, Clark Peter. Explaining Answers with Entailment Trees // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 7358–7370.

Damonte Marco, Monti Emilio. One Semantic Parser to Parse Them All: Sequence to Sequence Multi-Task Learning on Semantic Parsing Datasets // Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics. Online: Association for Computational Linguistics, VIII 2021. 173–184.

Dankers Verna, Bruni Elia, Hupkes Dieuwke. The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 4154–4175.

Dankers Verna, Langedijk Anna, McCurdy Kate, Williams Adina, Hupkes Dieuwke. Generalising to German Plural Noun Classes, from the Perspective of a Recurrent Neural Network // Proceedings of the 25th Conference on Computational Natural Language Learning. Online: Association for Computational Linguistics, XI 2021. 94–108.

Das Sarkar Snigdha Sarathi, Katiyar Arzoo, Passonneau Rebecca, Zhang Rui. CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 6338–6353.

Davis Forrest, Schijndel Marten van. Uncovering Constraint-Based Behavior in Neural Models via Targeted Fine-Tuning // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 1159–1171.

Dayanik Erenay, Padó Sebastian. Disentangling Document Topic and Author Gender in Multiple Languages: Lessons for Adversarial Debiasing // Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Online: Association for Computational Linguistics, IV 2021. 50–61.

Daza Angel, Frank Anette. X-SRL: A Parallel Cross-Lingual Semantic Role Labeling Dataset // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 3904–3914.

De Varda Andrea, Zamparelli Roberto. Multilingualism Encourages Recursion: a Transfer Study with mBERT // Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP. Seattle, Washington: Association for Computational Linguistics, VII 2022. 1–10.

Desai Shrey, Xu Jiacheng, Durrett Greg. Compressive Summarization with Plausibility and Salience Modeling // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 6259–6274.

Dessì Roberto, Baroni Marco. CNNs found to jump around more skillfully than RNNs: Compositional Generalization in Seq2seq Convolutional Networks // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 3919–3923.

Dhar Prajit, Plas Lonneke van der. Learning to Predict Novel Noun-Noun Compounds // Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). Florence, Italy: Association for Computational Linguistics, VIII 2019. 30–39.

Di Giovanni Marco, Brambilla Marco. Content-based Stance Classification of Tweets about the 2020 Italian Constitutional Referendum // Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media. Online: Association for Computational Linguistics, VI 2021. 14–23.

Ding Ning, Xu Guangwei, Chen Yulin, Wang Xiaobin, Han Xu, Xie Pengjun, Zheng Haitao, Liu Zhiyuan. Few-NERD: A Few-shot Named Entity Recognition Dataset // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 3198–3213.

Dixon Lucas, Li John, Sorensen Jeffrey, Thain Nithum, Vasserman Lucy. Measuring and Mitigating Unintended Bias in Text Classification // Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery, 2018. 67–73. (AIES '18).

Dong Xin, Melo Gerard de. A Helping Hand: Transfer Learning for Deep Sentiment Analysis // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, VII 2018. 2524–2534.

Dou Zi-Yi, Wang Xinyi, Hu Junjie, Neubig Graham. Domain Differential Adaptation for Neural Machine Translation // Proceedings of the 3rd Workshop on Neural Generation and Translation. Hong Kong: Association for Computational Linguistics, XI 2019. 59–69.

Douka Stella, Abdine Hadi, Vazirgiannis Michalis, El Hamdani Rajaa, Restrepo Amariles David. JuriBERT: A Masked-Language Model Adaptation for French Legal Text // Proceedings of the Natural Legal Language Processing Workshop 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 95–101.

Du Jingfei, Ott Myle, Li Haoran, Zhou Xing, Stoyanov Veselin. General Purpose Text Embeddings from Pre-trained Language Models for Scalable Inference // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020a. 3018–3030.

Du Mengnan, Manjunatha Varun, Jain Rajiv, Deshpande Ruchi, Dernoncourt Franck, Gu Jiuxiang, Sun Tong, Hu Xia. Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU models // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021a. 915–929.

Du Xinya, He Luheng, Li Qi, Yu Dian, Pasupat Panupong, Zhang Yuan. QA-Driven Zero-shot Slot Filling with Weak Supervision Pretraining // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021b. 654–664.

Du Yingjun, Holla Nithin, Zhen Xiantong, Snoek Cees, Shutova Ekaterina. Meta-Learning with Variational Semantic Memory for Word Sense Disambiguation // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021c. 5254–5268.

Du Yuheng, Oraby Shereen, Perera Vittorio, Shen Minmin, Narayan-Chen Anjali, Chung Tagyoung, Venkatesh Anushree, Hakkani-Tur Dilek. Schema-Guided Natural Language Generation // Proceedings of the 13th International Conference on Natural Language Generation. Dublin, Ireland: Association for Computational Linguistics, XII 2020b. 283–295.

Dubois Yann, Dagan Gautier, Hupkes Dieuwke, Bruni Elia. Location Attention for Extrapolation to Longer Sequences // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 403–413.

Dutta Subhabrata, Juneja Jeevesh, Das Dipankar, Chakraborty Tanmoy. Can Unsupervised Knowledge Transfer from Social Discussions Help Argument Mining? // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 7774–7786.

Eisape Tiwalayo, Zaslavsky Noga, Levy Roger. Cloze Distillation: Improving Neural Language Models with Human Next-Word Prediction // Proceedings of the 24th Conference on Computational Natural Language Learning. Online: Association for Computational Linguistics, XI 2020. 609–619.

Eisenberg Joshua, Finlayson Mark. A Simpler and More Generalizable Story Detector using Verb and Character Features // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, IX 2017. 2708–2715.

Ek Adam, Bernardy Jean-Philippe. Training Strategies for Neural Multilingual Morphological Inflection // Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. Online: Association for Computational Linguistics, VIII 2021. 260–267.

Ek Adam, Bernardy Jean-Philippe, Chatzikyriakidis Stergios. How does Punctuation Affect Neural Models in Natural Language Inference // Proceedings of the Probability and Meaning Conference (PaM 2020). Gothenburg: Association for Computational Linguistics, VI 2020. 109–116.

El Mekki Abdellah, El Mahdaouy Abdelkader, Berrada Ismail, Khoumsi Ahmed. Domain Adaptation for Arabic Cross-Domain and Cross-Dialect Sentiment Analysis from Contextualized Word Embedding // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 2824–2837.

ElJundi Obeida, Antoun Wissam, El Droubi Nour, Hajj Hazem, El-Hajj Wassim, Shaban Khaled. hULMonA: The Universal Language Model in Arabic // Proceedings of the Fourth Arabic Natural Language Processing Workshop. Florence, Italy: Association for Computational Linguistics, VIII 2019. 68–77.

Elangovan Aparna, He Jiayuan, Verspoor Karin. Memorization vs. Generalization : Quantifying Data Leakage in NLP Performance Evaluation // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, IV 2021. 1325–1335.

Falk Neele, Strakatova Yana, Huber Eva, Hinrichs Erhard. Automatic Classification of Attributes in German Adjective-Noun Phrases // Proceedings of the 14th International Conference on Computational Semantics (IWCS). Groningen, The Netherlands (online): Association for Computational Linguistics, VI 2021. 239–249.

Farag Youmna, Yannakoudakis Helen. Multi-Task Learning for Coherence Modeling // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 629–639.

Finegan-Dollak Catherine, Kummerfeld Jonathan K., Zhang Li, Ramanathan Karthik, Sadasivam Sesh, Zhang Rui, Radev Dragomir. Improving Text-to-SQL Evaluation Methodology // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, VII 2018. 351–360.

Fisch Adam, Talmor Alon, Jia Robin, Seo Minjoon, Choi Eunsol, Chen Danqi. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension // Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Hong Kong, China: Association for Computational Linguistics, XI 2019. 1–13.

FitzGerald Jack, Hench Christopher, Peris Charith, Mackie Scott, Rottmann Kay, Sanchez Ana, Nash Aaron, Urbach Liam, Kakarala Vishesh, Singh Richa, Ranganath Swetha, Crist Laurie, Britan Misha, Leeuwis Wouter, Tur Gokhan, Natarajan Prem. MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages. 2022.

Flachs Simon, Lacroix Ophélie, Yannakoudakis Helen, Rei Marek, Søgaard Anders. Grammatical Error Correction in Low Error Density Domains: A New Benchmark and Analyses // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 8467–8478.

Forbes Maxwell, Hwang Jena D., Shwartz Vered, Sap Maarten, Choi Yejin. Social Chemistry 101: Learning to Reason about Social and Moral Norms // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 653–670.

Frank Robert, Petty Jackson. Sequence-to-Sequence Networks Learn the Meaning of Reflexive Anaphora // Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference. Barcelona, Spain (online): Association for Computational Linguistics, XII 2020. 154–164.

Freitag Markus, Al-Onaizan Yaser. Fast Domain Adaptation for Neural Machine Translation. 2016.

Fried Daniel, Kitaev Nikita, Klein Dan. Cross-Domain Generalization of Neural Constituency Parsers // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 323–330.

Friedman Dan, Dodge Ben, Chen Danqi. Single-dataset Experts for Multi-dataset Question Answering // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 6128–6137.

Gai Yu, Jain Paras, Zhang Wendi, Gonzalez Joseph, Song Dawn, Stoica Ion. Grounded Graph Decoding improves Compositional Generalization in Question Answering // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 1829–1838.

Gan Yujian, Chen Xinyun, Purver Matthew. Exploring Underexplored Limitations of Cross-Domain Text-to-SQL Generalization // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 8926–8931.

Gao Shuyang, Agarwal Sanchit, Jin Di, Chung Tagyoung, Hakkani-Tur Dilek. From Machine Reading Comprehension to Dialogue State Tracking: Bridging the Gap // Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI. Online: Association for Computational Linguistics, VII 2020. 79–89.

Garcia-Silva Andres, Berrio Cristian, Gómez-Pérez José Manuel. An Empirical Study on Pre-trained Embeddings and Language Models for Bot Detection // Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). Florence, Italy: Association for Computational Linguistics, VIII 2019. 148–155.

Gardner Matt, Artzi Yoav, Basmov Victoria, Berant Jonathan, Bogin Ben, Chen Sihao, Dasigi Pradeep, Dua Dheeru, Elazar Yanai, Gottumukkala Ananth, Gupta Nitish, Hajishirzi Hannaneh, Ilharco Gabriel, Khashabi Daniel, Lin Kevin, Liu Jiangming, Liu Nelson F., Mulcaire Phoebe, Ning Qiang, Singh Sameer, Smith Noah A., Subramanian Sanjay, Tsarfaty Reut, Wallace Eric, Zhang Ally, Zhou Ben. Evaluating Models' Local Decision Boundaries via Contrast Sets // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 1307–1323.

Geiger Atticus, Cases Ignacio, Karttunen Lauri, Potts Chris. Posing fair generalization tasks for natural language inference // arXiv preprint arXiv:1911.00811. 2019a.

Geiger Atticus, Cases Ignacio, Karttunen Lauri, Potts Christopher. Posing Fair Generalization Tasks for Natural Language Inference // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019b. 4485–4495.

Geiger Atticus, Richardson Kyle, Potts Christopher. Neural Natural Language Inference Models Partially Embed Theories of Lexical Entailment and Negation // Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Online: Association for Computational Linguistics, XI 2020. 163–173.

Geng Ruiying, Li Binhua, Li Yongbin, Zhu Xiaodan, Jian Ping, Sun Jian. Induction Networks for Few-Shot Text Classification // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 3904–3913.

Ghazarian Sarik, Hedayatnia Behnam, Papangelis Alexandros, Liu Yang, Hakkani-Tur Dilek. What is wrong with you?: Leveraging User Sentiment for Automatic Dialog Evaluation // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 4194–4204.

Ghosh Sayan, Qi Zheng, Chaturvedi Snigdha, Srivastava Shashank. How Helpful is Inverse Reinforcement Learning for Table-to-Text Generation? // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021. 71–79.

Gillick Daniel, Kulkarni Sayali, Lansing Larry, Presta Alessandro, Baldridge Jason, Ie Eugene, Garcia-Olano Diego. Learning Dense Representations for Entity Retrieval // Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). Hong Kong, China: Association for Computational Linguistics, XI 2019. 528–537.

Glockner Max, Shwartz Vered, Goldberg Yoav. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics, VII 2018. 650–655.

Goldman Omer, Guriel David, Tsarfaty Reut. (Un)solving Morphological Inflection: Lemma Overlap Artificially Inflates Models' Performance // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 864–870.

Gontier Nicolas, Sinha Koustuv, Reddy Siva, Pal Chris. Measuring systematic generalization in neural proof generation with transformers // Advances in Neural Information Processing Systems. 2020. 33. 22231–22242.

Goodwin Emily, Reddy Siva, O'Donnell Timothy, Bahdanau Dzmitry. Compositional Generalization in Dependency Parsing // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 6482–6493.

Goodwin Emily, Sinha Koustuv, O'Donnell Timothy J. Probing Linguistic Systematicity // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 1958–1969.

Grivas Andreas, Alex Beatrice, Grover Claire, Tobin Richard, Whiteley William. Not a cute stroke: Analysis of Rule- and Neural Network-based Information Extraction Systems for Brain Radiology Reports // Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis. Online: Association for Computational Linguistics, XI 2020. 24–37.

Guan Jian, Feng Zhuoer, Chen Yamei, He Ruilin, Mao Xiaoxi, Fan Changjie, Huang Minlie. LOT: A Story-Centric Benchmark for Evaluating Chinese Long Text Understanding and Generation // Transactions of the Association for Computational Linguistics. 2022. 10. 434–451.

Guan Jian, Zhang Zhexin, Feng Zhuoer, Liu Zitao, Ding Wenbiao, Mao Xiaoxi, Fan Changjie, Huang Minlie. OpenMEVA: A Benchmark for Evaluating Open-ended Story Generation Metrics // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 6394–6407.

Gung James, Palmer Martha. Predicate Representations and Polysemy in VerbNet Semantic Parsing // Proceedings of the 14th International Conference on Computational Semantics (IWCS). Groningen, The Netherlands (online): Association for Computational Linguistics, VI 2021. 51–62.

Guo Demi, Kim Yoon, Rush Alexander. Sequence-Level Mixed Sample Data Augmentation // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 5547–5552.

Gupta Ashim, Srikumar Vivek. X-Fact: A New Benchmark Dataset for Multilingual Fact Checking // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021. 675–682.

Gupta Nitish, Lewis Mike. Neural Compositional Denotational Semantics for Question Answering // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, X-XI 2018. 2152–2161.

Gupta Nitish, Singh Sameer, Gardner Matt, Roth Dan. Paired Examples as Indirect Supervision in Latent Decision Models // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 5774–5785.

Guriel David, Goldman Omer, Tsarfaty Reut. Morphological Reinflection with Multiple Arguments: An Extended Annotation schema and a Georgian Case Study // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 196–202.

Hahn Vanessa, Ruiter Dana, Kleinbauer Thomas, Klakow Dietrich. Modeling Profanity and Hate Speech in Social Media with Semantic Subspaces // Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021). Online: Association for Computational Linguistics, VIII 2021. 6–16.

Haley Coleman. This is a BERT. Now there are several of them. Can they generalize to novel words? // Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Online: Association for Computational Linguistics, XI 2020. 333–341.

Hallmann Koen, Kunneman Florian, Liebrecht Christine, Bosch Antal van den, Mulken Margot van. Sarcastic Soulmates: Intimacy and irony markers in social media messaging // Linguistic Issues in Language Technology, Volume 14, 2016 - Modality: Logic, Semantics, Annotation, and Machine Learning. sept 2016.

Han Mingyue, Wang Yinglin. Doing Good or Doing Right? Exploring the Weakness of Commonsense Causal Reasoning Models // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021. 151–157.

Han Xing, Lundin Jessica. Multi-Pair Text Style Transfer for Unbalanced Data via Task-Adaptive Meta-Learning // Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing. Online: Association for Computational Linguistics, VIII 2021. 28–35.

Han Xu, Gao Tianyu, Lin Yankai, Peng Hao, Yang Yaoliang, Xiao Chaojun, Liu Zhiyuan, Li Peng, Zhou Jie, Sun Maosong. More Data, More Relations, More Context and More Openness: A Review and Outlook for Relation Extraction // Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. Suzhou, China: Association for Computational Linguistics, XII 2020. 745–758.

Haneczok Jacek, Jacquet Guillaume, Piskorski Jakub, Stefanovitch Nicolas. Fine-grained Event Classification in News-like Text Snippets - Shared Task 2, CASE 2021 // Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021). Online: Association for Computational Linguistics, VIII 2021. 179–192.

Hanselowski Andreas, PVS Avinesh, Schiller Benjamin, Caspelherr Felix, Chaudhuri Debanjan, Meyer Christian M., Gurevych Iryna. A Retrospective Analysis of the Fake News Challenge Stance-Detection Task // Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics, VIII 2018. 1859–1874.

Hansen Victor Petrén Bach, Søgaard Anders. Guideline Bias in Wizard-of-Oz Dialogues // Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future. Online: Association for Computational Linguistics, VIII 2021. 8–14.

Harrigian Keith, Aguirre Carlos, Dredze Mark. Do Models of Mental Health Based on Social Media Data Generalize? // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 3774–3788.

Havrylov Serhii, Kruszewski Germán, Joulin Armand. Cooperative Learning of Disjoint Syntax and Semantics // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 1118–1128.

He Qi, Wang Han, Zhang Yue. Enhancing Generalization in Natural Language Inference by Syntax // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 4973–4978.

He Ruidan, Lee Wee Sun, Ng Hwee Tou, Dahlmeier Daniel. Adaptive Semi-supervised Learning for Cross-domain Sentiment Classification // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, X-XI 2018a. 3467–3476.

He Yu, Li Jianxin, Song Yangqiu, He Mutian, Peng Hao. Time-evolving Text Classification with Deep Neural Networks // Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. 7 2018b. 2241–2247.

Hendrycks Dan, Liu Xiaoyuan, Wallace Eric, Dziedzic Adam, Krishnan Rishabh, Song Dawn. Pretrained Transformers Improve Out-of-Distribution Robustness // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 2744–2751.

Hershcovich Daniel, Abend Omri, Rappoport Ari. Multitask Parsing Across Semantic Representations // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, VII 2018. 373–385.

Herzig Jonathan, Berant Jonathan. Decoupling Structure and Lexicon for Zero-Shot Semantic Parsing // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, X-XI 2018. 1619–1629.

Herzig Jonathan, Berant Jonathan. Span-based Semantic Parsing for Compositional Generalization // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 908–921.

Hessel Jack, Lee Lillian. Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 1648–1659.

Hettiarachchi Hansi, Ranasinghe Tharindu. TransWiC at SemEval-2021 Task 2: Transformer-based Multilingual and Cross-lingual Word-in-Context Disambiguation // Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). Online: Association for Computational Linguistics, VIII 2021. 771–779.

Hitschler Julian, Berg Esther van den, Rehbein Ines. Authorship Attribution with Convolutional Neural Networks and POS-Eliding // Proceedings of the Workshop on Stylistic Variation. Copenhagen, Denmark: Association for Computational Linguistics, IX 2017. 53–58.

Hoffmann Jordan, Borgeaud Sebastian, Mensch Arthur, Buchatskaya Elena, Cai Trevor, Rutherford Eliza, Casas Diego de Las, Hendricks Lisa Anne, Welbl Johannes, Clark Aidan, Hennigan Tom, Noland Eric, Millican Katie, Driessche George van den, Damoc Bogdan, Guy Aurelia, Osindero Simon, Simonyan Karen, Elsen Erich, Rae Jack W., Vinyals Oriol, Sifre Laurent. Training Compute-Optimal Large Language Models. 2022.

Hofmann Valentin, Pierrehumbert Janet, Schütze Hinrich. Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 3594–3608.

Hu Jennifer, Gauthier Jon, Qian Peng, Wilcox Ethan, Levy Roger. A Systematic Assessment of Syntactic Generalization in Neural Language Models // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020a. 1725–1744.

Hu Junjie, Ruder Sebastian, Siddhant Aditya, Neubig Graham, Firat Orhan, Johnson Melvin. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation // Proceedings of the 37th International Conference on Machine Learning. 119. 13–18 Jul 2020b. 4411–4421. (Proceedings of Machine Learning Research).

Hu Ziniu, Sun Yizhou, Chang Kai-Wei. Relation-Guided Pre-Training for Open-Domain Question Answering // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 3431–3448.

Hua Xinyu, Wang Lu. Efficient Argument Structure Extraction with Transfer Learning and Active Learning // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 423–437.

Huang Jiaxin, Li Chunyuan, Subudhi Krishan, Jose Damien, Balakrishnan Shobana, Chen Weizhu, Peng Baolin, Gao Jianfeng, Han Jiawei. Few-Shot Named Entity Recognition: An Empirical Baseline Study // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021a. 10408–10423.

Huang Kaiyu, Huang Degen, Liu Zhuang, Mo Fengran. A Joint Multiple Criteria Model in Transfer Learning for Cross-domain Chinese Word Segmentation // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020a. 3873–3882.

Huang Kuan-Hao, Ahmad Wasi, Peng Nanyun, Chang Kai-Wei. Improving Zero-Shot Cross-Lingual Transfer Learning via Robust Training // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021b. 1684–1697.

Huang William, Liu Haokun, Bowman Samuel R. Counterfactually-Augmented SNLI Training Data Does Not Yield Better Generalization Than Unaugmented Data // Proceedings of the First Workshop on Insights from Negative Results in NLP. Online: Association for Computational Linguistics, XI 2020b. 82–87.

Huang Yi, Feng Junlan, Wu Xiaoting, Du Xiaoyu. Counterfactual Matters: Intrinsic Probing For Dialogue State Tracking // The First Workshop on Evaluations and Assessments of Neural Conversation Systems. Online: Association for Computational Linguistics, XI 2021c. 1–6.

Huang Yufan, Zhang Yanzhe, Chen Jiaao, Wang Xuezhi, Yang Diyi. Continual Learning for Text Classification with Information Disentanglement Based Regularization // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021d. 2736–2746.

Huber Christian, Hussain Juan, Nguyen Tuan-Nam, Song Kaihang, Stüker Sebastian, Waibel Alexander. Supervised Adaptation of Sequence-to-Sequence Speech Recognition Systems using Batch-Weighting // Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems. Suzhou, China: Association for Computational Linguistics, XII 2020. 9–17.

Huo Siyu, Ma Tengfei, Chen Jie, Chang Maria, Wu Lingfei, Witbrock Michael. Graph Enhanced Cross-Domain Text-to-SQL Generation // Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13). Hong Kong: Association for Computational Linguistics, XI 2019. 159–163.

Hupkes Dieuwke, Sara , Zuidema Willem. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure // Journal of Artificial Intelligence Research. 2018. 61. 907–926.

Hupkes Dieuwke, Dankers Verna, Mul Mathijs, Bruni Elia. Compositionality Decomposed: How do Neural Networks Generalise? // Journal of Artificial Intellgence Research. 2020. 67. 757–795.

Jagfeld Glorianna, Jenne Sabrina, Vu Ngoc Thang. Sequence-to-Sequence Models for Data-to-Text Natural Language Generation: Word- vs. Character-based Processing and Output Diversity // Proceedings of the 11th International Conference on Natural Language Generation. Tilburg University, The Netherlands: Association for Computational Linguistics, XI 2018. 221–232.

Jambor Dora, Bahdanau Dzmitry. LAGr: Label Aligned Graphs for Better Systematic Generalization in Semantic Parsing // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 3295–3308.

Jhunjhunwala Megha, Bryant Caleb, Shah Pararth. Multi-Action Dialog Policy Learning with Interactive Human Teaching // Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 1st virtual meeting: Association for Computational Linguistics, VII 2020. 290–296.

Jiang Haoming, He Pengcheng, Chen Weizhu, Liu Xiaodong, Gao Jianfeng, Zhao Tuo. SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 2177–2190.

Jiang Nanjiang, Marneffe Marie-Catherine de. Do You Know That Florence Is Packed with Visitors? Evaluating State-of-the-art Models of Speaker Commitment // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 4208–4213.

Jiang Yichen, Bansal Mohit. Inducing Transformer's Compositional Generalization Ability via Auxiliary Sequence Prediction Tasks // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 6253–6265.

Jin Hailong, Dong Tiansi, Hou Lei, Li Juanzi, Chen Hui, Dai Zelin, Yincen Qu. How Can Cross-lingual Knowledge Contribute Better to Fine-Grained Entity Typing? // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 3071–3081.

Jin Xisen, Lin Bill Yuchen, Rostami Mohammad, Ren Xiang. Learn Continually, Generalize Rapidly: Lifelong Knowledge Accumulation for Few-shot Learning // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 714–729.

Joshi Mandar, Choi Eunsol, Levy Omer, Weld Daniel, Zettlemoyer Luke. pair2vec: Compositional Word-Pair Embeddings for Cross-Sentence Inference // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 3597–3608.

Joshi Nitish, He He. An Investigation of the (In)effectiveness of Counterfactually Augmented Data // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 3668–3681.

Joshi Pratik, Aditya Somak, Sathe Aalok, Choudhury Monojit. TaxiNLI: Taking a Ride up the NLU Hill // Proceedings of the 24th Conference on Computational Natural Language Learning. Online: Association for Computational Linguistics, XI 2020. 41–55.

Kachuee Mohammad, Yuan Hao, Kim Young-Bum, Lee Sungjin. Self-Supervised Contrastive Learning for Efficient User Satisfaction Prediction in Conversational Agents // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 4053–4064.

Kale Mihir, Rastogi Abhinav. Text-to-Text Pre-Training for Data-to-Text Tasks // Proceedings of the 13th International Conference on Natural Language Generation. Dublin, Ireland: Association for Computational Linguistics, XII 2020. 97–102.

Kalouli Aikaterini-Lida, Crouch Richard, Paiva Valeria de. Hy-NLI: a Hybrid system for Natural Language Inference // Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, XII 2020. 5235–5249.

Kanashiro Pereira Lis, Taya Yuki, Kobayashi Ichiro. Multi-Layer Random Perturbation Training for improving Model Generalization Efficiently // Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 303–310.

Kann Katharina, Bjerva Johannes, Augenstein Isabelle, Plank Barbara, Søgaard Anders. Character-level Supervision for Low-resource POS Tagging // Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP. Melbourne: Association for Computational Linguistics, VII 2018. 1–11.

Karan Mladen, Šnajder Jan. Cross-Domain Detection of Abusive Language Online // Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Brussels, Belgium: Association for Computational Linguistics, X 2018. 132–137.

Karimi Mahabadi Rabeeh, Belinkov Yonatan, Henderson James. End-to-End Bias Mitigation by Modelling Biases in Corpora // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 8706–8716.

Karimi Mahabadi Rabeeh, Ruder Sebastian, Dehghani Mostafa, Henderson James. Parameter-efficient Multi-task Fine-tuning for Transformers via Shared Hypernetworks // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 565–576.

Kaushik Divyansh, Hovy Eduard, Lipton Zachary. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data // International Conference on Learning Representations. 2019.

Kavumba Pride, Heinzerling Benjamin, Brassard Ana, Inui Kentaro. Learning to Learn to be Right for the Right Reasons // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 3890–3898.

Kavumba Pride, Takahashi Ryo, Oda Yusuke. Are Prompt-based Models Clueless? // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 2333–2352.

Kedia Akhil, Chinthakindi Sai Chetan. Keep Learning: Self-supervised Meta-learning for Learning from Inference // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, IV 2021. 63–77.

Keskar Nitish Shirish, McCann Bryan, Xiong Caiming, Socher Richard. The Thieves on Sesame Street are Polyglots - Extracting Multilingual Models from Monolingual APIs // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 6203–6207.

Keysers Daniel, Schärli Nathanael, Scales Nathan, Buisman Hylke, Furrer Daniel, Kashubin Sergii, Momchev Nikola, Sinopalnikov Danila, Stafiniak Lukasz, Tihon Tibor, others . Measuring Compositional Generalization: A Comprehensive Method on Realistic Data // International Conference on Learning Representations. 2019.

Khaddaj Alaa, Hajj Hazem, El-Hajj Wassim. Improved Generalization of Arabic Text Classifiers // Proceedings of the Fourth Arabic Natural Language Processing Workshop. Florence, Italy: Association for Computational Linguistics, VIII 2019. 167–174.

Khashabi Daniel, Min Sewon, Khot Tushar, Sabharwal Ashish, Tafjord Oyvind, Clark Peter, Hajishirzi Hannaneh. UNIFIEDQA: Crossing Format Boundaries with a Single QA System // Findings of the

Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 1896–1907.

Khishigsuren Temuulen, Bella Gábor, Batsuren Khuyagbaatar, Freihat Abed Alhakim, Nair Nandu Chandran, Ganbold Amarsanaa, Khalilia Hadi, Chandrashekar Yamini, Giunchiglia Fausto. Using Linguistic Typology to Enrich Multilingual Lexicons: the Case of Lexical Gaps in Kinship // CoRR. 2022. abs/2204.05049.

Khot Tushar, Richardson Kyle, Khashabi Daniel, Sabharwal Ashish. Hey AI, Can You Solve Complex Tasks by Talking to Agents? // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 1808–1823.

Kim Juyong, Ravikumar Pradeep, Ainslie Joshua, Ontanon Santiago. Improving Compositional Generalization in Classification Tasks via Structure Annotations // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021. 637–645.

Kim Najoung, Linzen Tal. COGS: A Compositional Generalization Challenge Based on Semantic Interpretation // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 9087–9105.

Kim Young-Bum, Stratos Karl, Kim Dongchan. Domain Attention with an Ensemble of Experts // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, VII 2017. 643–653.

Kimura Mayuko, Kanashiro Pereira Lis, Kobayashi Ichiro. Towards a Language Model for Temporal Commonsense Reasoning // Proceedings of the Student Research Workshop Associated with RANLP 2021. Online: INCOMA Ltd., IX 2021. 78–84.

King Milton, Cook Paul. Supervised and unsupervised approaches to measuring usage similarity // Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications. Valencia, Spain: Association for Computational Linguistics, IV 2017. 47–52.

Kobayashi Sosuke, Yokoi Sho, Suzuki Jun, Inui Kentaro. Efficient Estimation of Influence of a Training Instance // Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing. Online: Association for Computational Linguistics, XI 2020. 41–47.

Kodner Jordan, Khalifa Salam, Batsuren Khuyagbaatar, Dolatian Hossep, Cotterell Ryan, Akkus Faruk, Anastasopoulos Antonios, Andrushko Taras, Arora Aryaman, Atanalov Nona, Bella Gábor, Budianskaya Elena, Ghanggo Ate Yustinus, Goldman Omer, Guriel David, Guriel Simon, Guriel-Agiashvili Silvia, Kieraś Witold, Krizhanovsky Andrew, Krizhanovsky Natalia, Marchenko Igor, Markowska Magdalena, Mashkovtseva Polina, Nepomniashchaya Maria, Rodionova Daria, Scheifer Karina, Sorova Alexandra, Yemelina Anastasia, Young Jeremiah, Vylomova Ekaterina. SIGMORPHON–UniMorph 2022 Shared Task 0: Generalization and Typologically Diverse Morphological Inflection // Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. Seattle, Washington: Association for Computational Linguistics, VII 2022. 176–203.

Koh Pang Wei, Sagawa Shiori, Marklund Henrik, Xie Sang Michael, Zhang Marvin, Balsubramani Akshay, Hu Weihua, Yasunaga Michihiro, Phillips Richard Lanas, Gao Irena, others . Wilds: A benchmark of in-the-wild distribution shifts // International Conference on Machine Learning. 2021. 5637–5664.

Körner Erik, Wiedemann Gregor, Hakimi Ahmad Dawar, Heyer Gerhard, Potthast Martin. On Classifying whether Two Texts are on the Same Side of an Argument // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 10130–10138.

Korrel Kris, Hupkes Dieuwke, Dankers Verna, Bruni Elia. Transcoding Compositionally: Using Attention to Find More Generalizable Solutions // Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Florence, Italy: Association for Computational Linguistics, VIII 2019. 1–11.

Koto Fajri, Baldwin Timothy, Lau Jey Han. Cloze Evaluation for Deeper Understanding of Commonsense Stories in Indonesian // Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022). Dublin, Ireland: Association for Computational Linguistics, V 2022. 8–16.

Kouris Panagiotis, Alexandridis Georgios, Stafylopatis Andreas. Abstractive Text Summarization: Enhancing Sequence-to-Sequence Models Using Word Sense Disambiguation and Semantic Content Generalization // Computational Linguistics. XII 2021. 47, 4. 813–859.

Kreutzer Julia, Sokolov Artem, Riezler Stefan. Bandit Structured Prediction for Neural Sequence-to-Sequence Learning // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, VII 2017. 1503–1513.

Kumar Gaurav, Koehn Philipp, Khudanpur Sanjeev. Learning Feature Weights using Reward Modeling for Denoising Parallel Corpora // Proceedings of the Sixth Conference on Machine Translation. Online: Association for Computational Linguistics, XI 2021. 1100–1109.

Kumar Sachin, Wintner Shuly, Smith Noah A., Tsvetkov Yulia. Topics to Avoid: Demoting Latent Confounds in Text Classification // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019a. 4153–4163.

Kumar Sawan, Jat Sharmistha, Saxena Karan, Talukdar Partha. Zero-shot Word Sense Disambiguation using Sense Definition Embeddings // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019b. 5670–5681.

Kumar Vinayshekhar, Kumar Vaibhav, Bhutani Mukul, Rudnicky Alexander. An Empirical study to understand the Compositional Prowess of Neural Dialog Models // Proceedings of the Third Workshop on Insights from Negative Results in NLP. Dublin, Ireland: Association for Computational Linguistics, V 2022. 154–158.

Kunchukuttan Anoop, Khapra Mitesh, Singh Gurneet, Bhattacharyya Pushpak. Leveraging Orthographic Similarity for Multilingual Neural Transliteration // Transactions of the Association for Computational Linguistics. 2018. 6. 303–316.

Lake Brenden, Baroni Marco. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks // proceedings of the 35th International Conference on Machine Learning (ICML). 2018. 4487–4499.

Lakretz Yair, Desbordes Theo, Hupkes Dieuwke, Dehaene Stanislas. Causal Transformers Perform Below Chance on Recursive Nested Constructions, Unlike Humans // CoRR. 2021a. abs/2110.07240.

Lakretz Yair, Hupkes Dieuwke, Vergallito Alessandra, Marelli Marco, Baroni Marco, Dehaene Stanislas. Mechanisms for handling nested dependencies in neural-network language models and humans // Cognition. 2021b. 213. 104699. Special Issue in Honour of Jacques Mehler, Cognition's founding editor.

Lasri Karim, Lenci Alessandro, Poibeau Thierry. Does BERT really agree ? Fine-grained Analysis of Lexical Dependence on a Syntactic Task // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 2309–2315.

Lazaridou Angeliki, Kuncoro Adhi, Gribovskaya Elena, Agrawal Devang, Liska Adam, Terzi Tayfun, Gimenez Mai, Masson d'Autume Cyprien de, Kocisky Tomas, Ruder Sebastian, others . Mind the gap: Assessing temporal generalization in neural language models // Advances in Neural Information Processing Systems. 2021. 34. 29348–29363.

Lee Chia-Hsuan, Polozov Oleksandr, Richardson Matthew. KaggleDBQA: Realistic Evaluation of Text-to-SQL Parsers // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021a. 2261–2273.

Lee Seanie, Kang Minki, Lee Juho, Hwang Sung Ju. Learning to Perturb Word Embeddings for Out-of-distribution QA // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021b. 5583–5595.

Lee Seanie, Kim Donggyu, Park Jangwon. Domain-agnostic Question-Answering with Adversarial Training // Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Hong Kong, China: Association for Computational Linguistics, XI 2019. 196–202.

Lee Yohan. Improving End-to-End Task-Oriented Dialog System with A Simple Auxiliary Task // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 1296–1303.

Lent Heather, Yavuz Semih, Yu Tao, Niu Tong, Zhou Yingbo, Radev Dragomir, Lin Xi Victoria. Testing Cross-Database Semantic Parsers With Canonical Utterances // Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 73–83.

Lepori Michael, Linzen Tal, McCoy R. Thomas. Representations of Syntax [MASK] Useful: Effects of Constituency and Dependency Structure in Recursive LSTMs // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 3306–3316.

Levy Omer, Seo Minjoon, Choi Eunsol, Zettlemoyer Luke. Zero-Shot Relation Extraction via Reading Comprehension // Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada: Association for Computational Linguistics, VIII 2017. 333–342.

Lewis Patrick, Stenetorp Pontus, Riedel Sebastian. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, IV 2021. 1000–1008.

Li Belinda, Yu Jane, Khabsa Madian, Zettlemoyer Luke, Halevy Alon, Andreas Jacob. Quantifying Adaptability in Pre-trained Language Models with 500 Tasks // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, VII 2022. 4696–4715.

Li Chong, Zhang Cenyuan, Zheng Xiaoqing, Huang Xuanjing. Exploration and Exploitation: Two Ways to Improve Chinese Spelling Correction Models // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021a. 441–446.

Li Chunyuan, Gao Xiang, Li Yuan, Peng Baolin, Li Xiujun, Zhang Yizhe, Gao Jianfeng. Optimus: Organizing Sentences via Pre-trained Modeling of a Latent Space // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020a. 4678–4699.

Li Fayuan, Peng Weihua, Chen Yuguang, Wang Quan, Pan Lu, Lyu Yajuan, Zhu Yong. Event Extraction as Multi-turn Question Answering // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020b. 829–838.

Li Haonan, Vasardani Maria, Tomko Martin, Baldwin Timothy. Target Word Masking for Location Metonymy Resolution // Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, XII 2020c. 3696–3707.

Li Jane S.Y., Wilson Colin. Leveraging Paradigmatic Information in Inflection Acceptability Prediction: The JHU-SFU Submission to SIGMORPHON Shared Task 0.2 // Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. Online: Association for Computational Linguistics, VIII 2021. 289–294.

Li Tianda, Rashid Ahmad, Jafari Aref, Sharma Pranav, Ghodsi Ali, Rezagholizadeh Mehdi. How to Select One Among All ? An Empirical Study Towards the Robustness of Knowledge Distillation in Natural Language Understanding // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021b. 750–762.

Li Yafu, Yin Yongjing, Chen Yulong, Zhang Yue. On Compositional Generalization of Neural Machine Translation // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021c. 4767–4780.

Li Yitong, Baldwin Timothy, Cohn Trevor. What's in a Domain? Learning Domain-Robust Text Representations using Adversarial Training // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics, VI 2018. 474–479.

Li Yuanpeng, Zhao Liang, Wang Jianyu, Hestness Joel. Compositional Generalization for Primitive Substitutions // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 4293–4302.

Li Yue, Zhang Jiong. Semi-supervised Meta-learning for Cross-domain Few-shot Intent Classification // Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing. Online: Association for Computational Linguistics, VIII 2021. 67–75.

Liang Chen, Zuo Simiao, Chen Minshuo, Jiang Haoming, Liu Xiaodong, He Pengcheng, Zhao Tuo, Chen Weizhu. Super Tickets in Pre-Trained Language Models: From Model Compression to Improving Generalization // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 6524–6538.

Liang Guanqing, Leung Cane Wing-Ki. Improving Model Generalization: A Chinese Named Entity Recognition Case Study // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021. 992–997.

Libovický Jindřich, Schmid Helmut, Fraser Alexander. Why don't people use character-level machine translation? // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 2470–2485.

Limisiewicz Tomasz, Mareček David, Rosa Rudolf. Universal Dependencies According to BERT: Both More Specific and More General // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 2710–2722.

Lin Bill Yuchen, Lee Dong-Ho, Shen Ming, Moreno Ryan, Huang Xiao, Shiralkar Prashant, Ren Xiang. TriggerNER: Learning with Entity Triggers as Explanations for Named Entity Recognition // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020a. 8503–8511.

Lin Bill Yuchen, Zhou Wangchunshu, Shen Ming, Zhou Pei, Bhagavatula Chandra, Choi Yejin, Ren Xiang. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020b. 1823–1840.

Lin Hsien-chin, Lubis Nurul, Hu Songbo, Niekerk Carel van, Geishauser Christian, Heck Michael, Feng Shutong, Gasic Milica. Domain-independent User Simulation with Transformers for Task-oriented Dialogue Systems // Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. Singapore and Online: Association for Computational Linguistics, VII 2021a. 445–456.

Lin Xi Victoria, Mihaylov Todor, Artetxe Mikel, Wang Tianlu, Chen Shuohui, Simig Daniel, Ott Myle, Goyal Naman, Bhosale Shruti, Du Jingfei, Pasunuru Ramakanth, Shleifer Sam, Koura Punit Singh, Chaudhary Vishrav, O'Horo Brian, Wang Jeff, Zettlemoyer Luke, Kozareva Zornitsa, Diab Mona, Stoyanov Veselin, Li Xian. Few-shot Learning with Multilingual Language Models. 2021b.

Lin Zehui, Wu Liwei, Wang Mingxuan, Li Lei. Learning Language Specific Sub-network for Multilingual Machine Translation // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021c. 293–305.

Lin Zhaojiang, Liu Bing, Madotto Andrea, Moon Seungwhan, Zhou Zhenpeng, Crook Paul, Wang Zhiguang, Yu Zhou, Cho Eunjoon, Subba Rajen, Fung Pascale. Zero-Shot Dialogue State Tracking via Cross-Task Transfer // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021d. 7890–7900.

Liu Fenglin, Gao Meng, Liu Yuanxin, Lei Kai. Self-Adaptive Scaling for Learnable Residual Structure // Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). Hong Kong, China: Association for Computational Linguistics, XI 2019a. 862–870.

Liu Linlin, Ding Bosheng, Bing Lidong, Joty Shafiq, Si Luo, Miao Chunyan. MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021a. 5834–5846.

Liu Linqing, Lewis Patrick S. H., Riedel Sebastian, Stenetorp Pontus. Challenges in Generalization in Open Domain Question Answering // CoRR. 2021b. abs/2109.01156.

Liu Miaofeng, Song Yan, Zou Hongbin, Zhang Tong. Reinforced Training Data Selection for Domain Adaptation // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019b. 1957–1968.

Liu Nelson F., Hershcovich Daniel, Kranzlein Michael, Schneider Nathan. Lexical Semantic Recognition // Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021). Online: Association for Computational Linguistics, VIII 2021c. 49–56.

Liu Qin, Zheng Rui, Rong Bao, Liu Jingyi, Liu ZhiHua, Cheng Zhanzhan, Qiao Liang, Gui Tao, Zhang Qi, Huang Xuanjing. Flooding-X: Improving BERT's Resistance to Adversarial Attacks via Loss-Restricted Fine-Tuning // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022a. 5634–5644.

Liu Yinhan, Ott Myle, Goyal Naman, Du Jingfei, Joshi Mandar, Chen Danqi, Levy Omer, Lewis Mike, Zettlemoyer Luke, Stoyanov Veselin. RoBERTa: A Robustly Optimized BERT Pretraining Approach // CoRR. 2019c. abs/1907.11692.

Liu Zheng, Zhang Wei, Chen Yan, Sun Weiyi, Du Tianchuan, Schroeder Benjamin. Towards Generalizeable Semantic Product Search by Text Similarity Pre-training on Search Click Logs // Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5). Dublin, Ireland: Association for Computational Linguistics, V 2022b. 224–233.

Liu Zhengyuan, Chen Nancy. Improving Multi-Party Dialogue Discourse Parsing via Domain Integration // Proceedings of the 2nd Workshop on Computational Approaches to Discourse. Punta Cana, Dominican Republic and Online: Association for Computational Linguistics, XI 2021. 122–127.

Liu Zhengyuan, Shi Ke, Chen Nancy. DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Parsing // Proceedings of the 2nd Workshop on Computational Approaches to Discourse. Punta Cana, Dominican Republic and Online: Association for Computational Linguistics, XI 2021d. 154–164.

Liu Zhongkun, Ren Pengjie, Chen Zhumin, Ren Zhaochun, Rijke Maarten de, Zhou Ming. Learning to Ask Conversational Questions by Optimizing Levenshtein Distance // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021e. 5638–5650.

Liu Zihan, Patwary Mostofa, Prenger Ryan, Prabhumoye Shrimai, Ping Wei, Shoeybi Mohammad, Catanzaro Bryan. Multi-Stage Prompting for Knowledgeable Dialogue Generation // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022c. 1317–1337.

Liu Zihan, Winata Genta Indra, Xu Peng, Fung Pascale. X2Parser: Cross-Lingual and Cross-Domain Framework for Task-Oriented Compositional Semantic Parsing // Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021). Online: Association for Computational Linguistics, VIII 2021f. 112–127.

Liu Zoey, Prud'hommeaux Emily. Data-driven Model Generalizability in Crosslinguistic Low-resource Morphological Segmentation // Transactions of the Association for Computational Linguistics. 2022. 10. 393–413.

Liška Adam, Kruszewski Germán, Baroni Marco. Memorize or generalize? Searching for a compositional RNN in a haystack // Proceedings of AEGAP (FAIM Joint Workshop on Architectures and Evaluation for Generality, Autonomy and Progress in AI). 2018.

Logeswaran Lajanugen, Chang Ming-Wei, Lee Kenton, Toutanova Kristina, Devlin Jacob, Lee Honglak. Zero-Shot Entity Linking by Reading Entity Descriptions // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 3449–3460.

Longpre Shayne, Lu Yi, DuBois Chris. On the Transferability of Minimal Prediction Preserving Inputs in Question Answering // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021a. 1288–1300.

Longpre Shayne, Perisetla Kartik, Chen Anthony, Ramesh Nikhil, DuBois Chris, Singh Sameer. Entity-Based Knowledge Conflicts in Question Answering // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021b. 7052–7063.

Loula João, Baroni Marco, Lake Brenden. Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks // Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium: Association for Computational Linguistics, XI 2018. 108–114.

Lu Zhichu, Arabshahi Forough, Labutov Igor, Mitchell Tom. Look-up and Adapt: A One-shot Semantic Parser // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 1129–1139.

Ludwig Florian, Dolos Klara, Zesch Torsten, Hobley Eleanor. Improving Generalization of Hate Speech Detection Systems to Novel Target Groups via Domain Adaptation // Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH). Seattle, Washington (Hybrid): Association for Computational Linguistics, VII 2022. 29–39.

Ma Kaixin, Ilievski Filip, Francis Jonathan, Ozaki Satoru, Nyberg Eric, Oltramari Alessandro. Exploring Strategies for Generalizable Commonsense Reasoning with Pre-trained Models // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 5474–5483.

Ma Yubo, Wang Zehao, Cao Yixin, Li Mukai, Chen Meiqi, Wang Kun, Shao Jing. Prompt for Extraction? PAIE: Prompting Argument Interaction for Event Argument Extraction // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 6759–6774.

Maharana Adyasha, Bansal Mohit. Adversarial Augmentation Policy Search for Domain and Cross-Lingual Generalization in Reading Comprehension // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 3723–3738.

Mahurkar Siddhant, Patil Rajaswa. LRG at SemEval-2020 Task 7: Assessing the Ability of BERT and Derivative Models to Perform Short-Edits Based Humor Grading // Proceedings of the Fourteenth Workshop on Semantic Evaluation. Barcelona (online): International Committee for Computational Linguistics, XII 2020. 858–864.

Malinin Andrey, Band Neil, Gal Yarin, Gales Mark J. F., Ganshin Alexander, Chesnokov German, Noskov Alexey, Ploskonosov Andrey, Prokhorenkova Liudmila, Provilkov Ivan, Raina Vatsal, Raina Vyas, Roginskiy Denis, Shmatova Mariya, Tigas Panagiotis, Yangel Boris. Shifts: A Dataset of Real Distributional Shift Across Multiple Large-Scale Tasks // Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual. 2021.

Malkiel Itzik, Wolf Lior. Maximal Multiverse Learning for Promoting Cross-Task Generalization of Fine-Tuned Language Models // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, IV 2021. 187–199.

Mao Yuren, Wang Zekai, Liu Weiwei, Lin Xuemin, Hu Wenbin. BanditMTL: Bandit-based Multi-task Learning for Text Classification // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 5506–5516.

Marasović Ana, Zhou Mengfei, Palmer Alexis, Frank Anette. Modal Sense Classification At Large: Paraphrase-Driven Sense Projection, Semantically Enriched Classification Models and Cross-Genre Evaluations // Linguistic Issues in Language Technology, Volume 14, 2016 - Modality: Logic, Semantics, Annotation, and Machine Learning. sept 2016.

Maronikolakis Antonis, Schütze Hinrich. Multidomain Pretrained Language Models for Green NLP // Proceedings of the Second Workshop on Domain Adaptation for NLP. Kyiv, Ukraine: Association for Computational Linguistics, IV 2021. 1–8.

Marzinotto Gabriel, Damnati Géraldine, Béchet Frédéric, Favre Benoît. Robust Semantic Parsing with Adversarial Learning for Domain Generalization // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 166–173.

McCann Bryan, Keskar Nitish Shirish, Xiong Caiming, Socher Richard. The natural language decathlon: Multitask learning as question answering // arXiv preprint arXiv:1806.08730. 2018.

McCoy R. Thomas, Frank Robert, Linzen Tal. Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks // Transactions of the Association for Computational Linguistics. 2020a. 8. 125–140.

McCoy R. Thomas, Min Junghyun, Linzen Tal. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance // Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Online: Association for Computational Linguistics, XI 2020b. 217–227.

McCoy Tom, Pavlick Ellie, Linzen Tal. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 3428–3448.

McCurdy Kate, Goldwater Sharon, Lopez Adam. Inflecting When There's No Majority: Limitations of Encoder-Decoder Neural Networks as Cognitive Models for German Plurals // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 1745–1756.

McHardy Robert, Adel Heike, Klinger Roman. Adversarial Training for Satire Detection: Controlling for Confounding Variables // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 660–665.

Mehta Sanket Vaibhav, Rao Jinfeng, Tay Yi, Kale Mihir, Parikh Ankur, Strubell Emma. Improving Compositional Generalization with Self-Training for Data-to-Text Generation // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 4205–4219.

Merity Stephen, Xiong Caiming, Bradbury James, Socher Richard. Pointer Sentinel Mixture Models // 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. 2017.

M'hamdi Meryem, Kim Doo Soon, Dernoncourt Franck, Bui Trung, Ren Xiang, May Jonathan. X-METRA-ADA: Cross-lingual Meta-Transfer learning Adaptation to Natural Language Understanding and Question Answering // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 3617–3632.

Mihaylov Todor, Frank Anette. Discourse-Aware Semantic Self-Attention for Narrative Reading Comprehension // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 2541–2552.

Min Junghyun, McCoy R. Thomas, Das Dipanjan, Pitler Emily, Linzen Tal. Syntactic Data Augmentation Increases Robustness to Inference Heuristics // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 2339–2352.

Min Sewon, Lewis Mike, Hajishirzi Hannaneh, Zettlemoyer Luke. Noisy Channel Language Model Prompting for Few-Shot Text Classification // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 5316–5330.

Mishra Swaroop, Khashabi Daniel, Baral Chitta, Hajishirzi Hannaneh. Cross-Task Generalization via Natural Language Crowdsourcing Instructions // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 3470–3487.

Mishra Swaroop, Sachdeva Bhavdeep Singh. Do We Need to Create Big Datasets to Learn a Task? // Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing. Online: Association for Computational Linguistics, XI 2020. 169–173.

Moeller Sarah, Kazeminejad Ghazaleh, Cowell Andrew, Hulden Mans. A Neural Morphological Analyzer for Arapaho Verbs Learned from a Finite State Transducer // Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages. Santa Fe, New Mexico, USA: Association for Computational Linguistics, VIII 2018. 12–20.

Moghimifar Farhad, Qu Lizhen, Zhuo Terry Yue, Haffari Gholamreza, Baktashmotlagh Mahsa. Neural-Symbolic Commonsense Reasoner with Relation Predictors // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021. 797–802.

Moosavi Nafise Sadat, Strube Michael. Lexical Features in Coreference Resolution: To be Used With Caution // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada: Association for Computational Linguistics, VII 2017. 14–19.

Moosavi Nafise Sadat, Strube Michael. Using Linguistic Features to Improve the Generalization Capability of Neural Coreference Resolvers // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, X-XI 2018. 193–203.

Mosca Edoardo, Agarwal Shreyash, Rando Ramírez Javier, Groh Georg. "That Is a Suspicious Reaction!": Interpreting Logits Variation to Detect NLP Adversarial Attacks // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 7806–7816.

Mueller Aaron, Frank Robert, Linzen Tal, Wang Luheng, Schuster Sebastian. Coloring the Blank Slate: Pre-training Imparts a Hierarchical Inductive Bias to Sequence-to-sequence Models // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 1352–1368.

Mul Mathijs, Zuidema Willem. Siamese recurrent networks learn first-order logic reasoning and exhibit zero-shot compositional generalization // CoRR, abs/1906.00180. 2019.

Muller Benjamin, Soldaini Luca, Koncel-Kedziorski Rik, Lind Eric, Moschitti Alessandro. Cross-Lingual GenQA: Open-Domain Question Answering with Answer Sentence Generation. 2021.

Nadeem Farah, Ostendorf Mari. Estimating Linguistic Complexity for Science Texts // Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications. New Orleans, Louisiana: Association for Computational Linguistics, VI 2018. 45–55.

Naderi Nona, Hirst Graeme. Using context to identify the language of face-saving // Proceedings of the 5th Workshop on Argument Mining. Brussels, Belgium: Association for Computational Linguistics, XI 2018. 111–120.

Naik Aakanksha, Rose Carolyn. Towards Open Domain Event Trigger Identification using Adversarial Domain Adaptation // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 7618–7624.

Nan Linyong, Radev Dragomir, Zhang Rui, Rau Amrit, Sivaprasad Abhinand, Hsieh Chiachun, Tang Xiangru, Vyas Aadit, Verma Neha, Krishna Pranav, Liu Yangxiaokang, Irwanto Nadia, Pan Jessica, Rahman Faiaz, Zaidi Ahmad, Mutuma Mutethia, Tarabar Yasin, Gupta Ankit, Yu Tao, Tan Yi Chern, Lin Xi Victoria, Xiong Caiming, Socher Richard, Rajani Nazneen Fatema. DART: Open-Domain

Structured Data Record to Text Generation // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 432–447.

Nangia Nikita, Bowman Samuel R. Human vs. Muppet: A Conservative Estimate of Human Performance on the GLUE Benchmark // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 4566–4575.

Nejadgholi Isar, Kiritchenko Svetlana. On Cross-Dataset Generalization in Automatic Detection of Online Abuse // Proceedings of the Fourth Workshop on Online Abuse and Harms. Online: Association for Computational Linguistics, XI 2020. 173–183.

Newman Benjamin, Hewitt John, Liang Percy, Manning Christopher D. The EOS Decision and Length Extrapolation // Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Online: Association for Computational Linguistics, XI 2020. 276–291.

Ng Nathan, Cho Kyunghyun, Ghassemi Marzyeh. SSMBA: Self-Supervised Manifold Based Data Augmentation for Improving Out-of-Domain Robustness // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 1268–1283.

Nguyen Thong, Yates Andrew, Zirikly Ayah, Desmet Bart, Cohan Arman. Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 8446–8459.

Nguyen Vincent, Karimi Sarvnaz, Xing Zhenchang. Combining Shallow and Deep Representations for Text-Pair Classification // Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association. Online: Australasian Language Technology Association, XII 2021. 68–78.

Nicolai Garrett, Silfverberg Miikka. Noise Isn't Always Negative: Countering Exposure Bias in Sequence-to-Sequence Inflection Models // Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, XII 2020. 2837–2846.

Nie Yixin, Williamson Mary, Bansal Mohit, Kiela Douwe, Weston Jason. I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 1699–1713.

Nye Maxwell, Solar-Lezama Armando, Tenenbaum Josh, Lake Brenden M. Learning compositional rules via neural program synthesis // Advances in Neural Information Processing Systems. 2020. 33. 10832–10842.

Ontanon Santiago, Ainslie Joshua, Fisher Zachary, Cvicek Vaclav. Making Transformers Solve Compositional Tasks // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 3591–3607.

Oren Inbar, Herzig Jonathan, Berant Jonathan. Finding needles in a haystack: Sampling Structurally-diverse Training Sets from Synthetic Data for Compositional Generalization // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 10793–10809.

Oren Inbar, Herzig Jonathan, Gupta Nitish, Gardner Matt, Berant Jonathan. Improving Compositional Generalization in Semantic Parsing // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 2482–2495.

Panda Subhadarshi, Levitan Sarah Ita. Detecting Multilingual COVID-19 Misinformation on Social Media via Contextualized Embeddings // Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. Online: Association for Computational Linguistics, VI 2021. 125–129.

Papangelis Alexandros, Gopalakrishnan Karthik, Padmakumar Aishwarya, Kim Seokhwan, Tur Gokhan, Hakkani-Tur Dilek. Generative Conversational Networks // Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue. Singapore and Online: Association for Computational Linguistics, VII 2021. 111–120.

Pappas Nikolaos, Henderson James. GILE: A Generalized Input-Label Embedding for Text Classification // Transactions of the Association for Computational Linguistics. 2019. 7. 139–155.

Patel Arkil, Bhattamishra Satwik, Blunsom Phil, Goyal Navin. Revisiting the Compositional Generalization Abilities of Neural Sequence Models // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 424–434.

Pedinotti Paolo, Rambelli Giulia, Chersoni Emmanuele, Santus Enrico, Lenci Alessandro, Blache Philippe. Did the Cat Drink the Coffee? Challenging Transformers with Generalized Event Knowledge // Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics. Online: Association for Computational Linguistics, VIII 2021. 1–11.

Pelicon Andraž, Shekhar Ravi, Martinc Matej, Škrlj Blaž, Purver Matthew, Pollak Senja. Zero-shot Cross-lingual Content Filtering: Offensive Language and Hate Speech Detection // Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. Online: Association for Computational Linguistics, IV 2021. 30–34.

Peng Baolin, Li Chunyuan, Zhang Zhu, Zhu Chenguang, Li Jinchao, Gao Jianfeng. RADDLE: An Evaluation Benchmark and Analysis Platform for Robust Task-oriented Dialog Systems // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 4418–4429.

Peng Baolin, Zhu Chenguang, Li Chunyuan, Li Xiujun, Li Jinchao, Zeng Michael, Gao Jianfeng. Few-shot Natural Language Generation for Task-Oriented Dialog // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 172–182.

Peng Nanyun, Dredze Mark. Multi-task Domain Adaptation for Sequence Tagging // Proceedings of the 2nd Workshop on Representation Learning for NLP. Vancouver, Canada: Association for Computational Linguistics, VIII 2017. 91–100.

Perez Ethan, Karamcheti Siddharth, Fergus Rob, Weston Jason, Kiela Douwe, Cho Kyunghyun. Finding Generalizable Evidence by Learning to Convince Q&A Models // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 2402–2411.

Perez Ethan, Kiela Douwe, Cho Kyunghyun. True Few-Shot Learning with Language Models // Advances in Neural Information Processing Systems. 2021.

Pérez-Mayos Laura, Ballesteros Miguel, Wanner Leo. How much pretraining data do language models need to learn syntax? // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 1571–1582.

Pham MinhQuang, Crego Josep, Yvon François, Senellart Jean. Generic and Specialized Word Embeddings for Multi-Domain Machine Translation // Proceedings of the 16th International Conference on Spoken Language Translation. Hong Kong: Association for Computational Linguistics, XI 2-3 2019.

Phang Jason, Févry Thibault, Bowman Samuel R. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks // ArXiv. 2018. abs/1811.01088.

Philip Jerin, Berard Alexandre, Gallé Matthias, Besacier Laurent. Monolingual Adapters for Zero-Shot Neural Machine Translation // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 4465–4470.

Phung Duy, Minh Tran Hieu, Nguyen Minh Van, Nguyen Thien Huu. Learning Cross-lingual Representations for Event Coreference Resolution with Multi-view Alignment and Optimal Transport // Proceedings of the 1st Workshop on Multilingual Representation Learning. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 62–73.

Picco Gabriele, Hoang Thanh Lam, Sbodio Marco Luca, Lopez Vanessa. Neural Unification for Logic Reasoning over Natural Language // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 3939–3950.

Pimentel Tiago, Ryskina Maria, Mielke Sabrina J., Wu Shijie, Chodroff Eleanor, Leonard Brian, Nicolai Garrett, Ghanggo Ate Yustinus, Khalifa Salam, Habash Nizar, El-Khaissi Charbel, Goldman Omer, Gasser Michael, Lane William, Coler Matt, Oncevay Arturo, Montoya Samame Jaime Rafael, Silva Villegas Gema Celeste, Ek Adam, Bernardy Jean-Philippe, Shcherbakov Andrey, Bayyr-ool Aziyana, Sheifer Karina, Ganieva Sofya, Plugaryov Matvey, Klyachko Elena, Salehi Ali, Krizhanovsky Andrew, Krizhanovsky Natalia, Vania Clara, Ivanova Sardana, Salchak Aelita, Straughn Christopher, Liu Zoey, Washington Jonathan North, Ataman Duygu, Kieraś Witold, Woliński Marcin, Suhardijanto Totok, Stoehr Niklas, Nuriah Zahroh, Ratan Shyam, Tyers Francis M., Ponti Edoardo M., Aiton Grant, Hatcher Richard J., Prud'hommeaux Emily, Kumar Ritesh, Hulden Mans, Barta Botond, Lakatos Dorina, Szolnok Gábor, Ács Judit, Raj Mohit, Yarowsky David, Cotterell Ryan, Ambridge Ben, Vylomova Ekaterina. SIGMORPHON 2021 Shared Task on Morphological Reinflection: Generalization Across Languages // Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. Online: Association for Computational Linguistics, VIII 2021. 229–259.

Plank Barbara. What to do about non-standard (or non-canonical) language in NLP // arXiv preprint arXiv:1608.07836. 2016.

Ponti Edoardo M., Vulić Ivan, Cotterell Ryan, Parovic Marinela, Reichart Roi, Korhonen Anna. Parameter Space Factorization for Zero-Shot Learning across Tasks and Languages // Transactions of the Association for Computational Linguistics. 2021. 9. 410–428.

Ponti Edoardo Maria, Glavaš Goran, Majewska Olga, Liu Qianchu, Vulić Ivan, Korhonen Anna. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 2362–2376.

Pouran Ben Veyseh Amir, Dernoncourt Franck, Dou Dejing, Nguyen Thien Huu. Exploiting the Syntax-Model Consistency for Neural Relation Extraction // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 8021–8032.

Power Alethea, Burda Yuri, Edwards Harri, Babuschkin Igor, Misra Vedant. Grokking: Generalization beyond overfitting on small algorithmic datasets // ICLR MATH-AI Workshop. 2021.

Pradeep Ronak, Ma Xueguang, Nogueira Rodrigo, Lin Jimmy. Scientific Claim Verification with VerT5erini // Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis. online: Association for Computational Linguistics, IV 2021. 94–103.

Prange Jakob, Schneider Nathan, Srikumar Vivek. Supertagging the Long Tail with Tree-Structured Decoding of Complex Categories // Transactions of the Association for Computational Linguistics. 2021. 9. 243–260.

Prickett Brandon, Traylor Aaron, Pater Joe. Seq2Seq Models with Dropout can Learn Generalizable Reduplication // Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology. Brussels, Belgium: Association for Computational Linguistics, X 2018. 93–100.

Qian Jing, ElSherief Mai, Belding Elizabeth, Wang William Yang. Learning to Decipher Hate Symbols // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 3006–3015.

Qian Peng, Naseem Tahira, Levy Roger, Fernandez Astudillo Ramón. Structural Guidance for Transformer Language Models // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 3735–3745.

Qin Chengwei, Joty Shafiq. Continual Few-shot Relation Learning via Embedding Space Regularization and Data Augmentation // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 2776–2789.

R. Menon Rakesh, Ghosh Sayan, Srivastava Shashank. CLUES: A Benchmark for Learning Classifiers using Natural Language Explanations // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 6523–6546.

Rae Jack W., Borgeaud Sebastian, Cai Trevor, Millican Katie, Hoffmann Jordan, Song Francis, Aslanides John, Henderson Sarah, Ring Roman, Young Susannah, Rutherford Eliza, Hennigan Tom, Menick Jacob, Cassirer Albin, Powell Richard, Driessche George van den, Hendricks Lisa Anne, Rauh

Maribeth, Huang Po-Sen, Glaese Amelia, Welbl Johannes, Dathathri Sumanth, Huang Saffron, Uesato Jonathan, Mellor John, Higgins Irina, Creswell Antonia, McAleese Nat, Wu Amy, Elsen Erich, Jayakumar Siddhant, Buchatskaya Elena, Budden David, Sutherland Esme, Simonyan Karen, Paganini Michela, Sifre Laurent, Martens Lena, Li Xiang Lorraine, Kuncoro Adhiguna, Nematzadeh Aida, Gribovskaya Elena, Donato Domenic, Lazaridou Angeliki, Mensch Arthur, Lespiau Jean-Baptiste, Tsimpoukelli Maria, Grigorev Nikolai, Fritz Doug, Sottiaux Thibault, Pajarskas Mantas, Pohlen Toby, Gong Zhitao, Toyama Daniel, d'Autume Cyprien de Masson, Li Yujia, Terzi Tayfun, Mikulik Vladimir, Babuschkin Igor, Clark Aidan, Casas Diego de Las, Guy Aurelia, Jones Chris, Bradbury James, Johnson Matthew, Hechtman Blake, Weidinger Laura, Gabriel Iason, Isaac William, Lockhart Ed, Osindero Simon, Rimell Laura, Dyer Chris, Vinyals Oriol, Ayoub Kareem, Stanway Jeff, Bennett Lorrayne, Hassabis Demis, Kavukcuoglu Koray, Irving Geoffrey. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. 2021.

Raffel Colin, Shazeer Noam, Roberts Adam, Lee Katherine, Narang Sharan, Matena Michael, Zhou Yanqi, Li Wei, Liu Peter J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // Journal of Machine Learning Research. 2020. 21, 140. 1–67.

Ranasinghe Tharindu, Orasan Constantin, Mitkov Ruslan. An Exploratory Analysis of Multilingual Word-Level Quality Estimation with Cross-Lingual Transformers // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021. 434–440.

Raunak Vikas, Kumar Vaibhav, Metze Florian, Callan Jaimie. On Compositionality in Neural Machine Translation // NeurIPS 2019 Context and Compositionality in Biological and Artificial Neural Systems Workshop. 2019.

Ravichander Abhilasha, Hovy Eduard, Suleman Kaheer, Trischler Adam, Cheung Jackie Chi Kit. On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT // Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics. Barcelona, Spain (Online): Association for Computational Linguistics, XII 2020. 88–102.

Rawat Bhanu Pratap Singh, Weng Wei-Hung, Min So Yeon, Raghavan Preethi, Szolovits Peter. Entity-Enriched Neural Models for Clinical Question Answering // Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. Online: Association for Computational Linguistics, VII 2020. 112–122.

Ray Chowdhury Jishnu, Caragea Cornelia, Caragea Doina. Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Online: Association for Computational Linguistics, VII 2020. 292–298.

Razeghi Yasaman, Logan IV Robert L, Gardner Matt, Singh Sameer. Impact of pretraining term frequencies on few-shot reasoning // CoRR. 2022. abs/2202.07206.

Reed Lena, Oraby Shereen, Walker Marilyn. Can Neural Generators for Dialogue Learn Sentence Planning and Discourse Structuring? // Proceedings of the 11th International Conference on Natural Language Generation. Tilburg University, The Netherlands: Association for Computational Linguistics, XI 2018. 284–295.

Ren Shuhuai, Zhang Jinchao, Li Lei, Sun Xu, Zhou Jie. Text AutoAugment: Learning Compositional Augmentation Policy for Text Classification // Proceedings of the 2021 Conference on Empirical

Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021a. 9029–9043.

Ren Xiaoying, Jiang Jing, Serena Khoo Ling Min, Chieu Hai Leong. Cross-Topic Rumor Detection using Topic-Mixtures // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, IV 2021b. 1534–1538.

Reuel Ann-Katrin, Peralta Sebastian, Sedoc João, Sherman Garrick, Ungar Lyle. Measuring the Language of Self-Disclosure across Corpora // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 1035–1047.

Ribeiro Marco Tulio, Wu Tongshuang, Guestrin Carlos, Singh Sameer. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 4902–4912.

Risch Julian, Krestel Ralf. Aggression Identification Using Deep Learning and Data Augmentation // Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). Santa Fe, New Mexico, USA: Association for Computational Linguistics, VIII 2018. 150–158.

Rivera-Soto Rafael A., Miano Olivia Elizabeth, Ordonez Juanita, Chen Barry Y., Khan Aleem, Bishop Marcus, Andrews Nicholas. Learning Universal Authorship Representations // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 913–919.

Robertson Alexander, Goldwater Sharon. Evaluating Historical Text Normalization Systems: How Well Do They Generalize? // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). New Orleans, Louisiana: Association for Computational Linguistics, VI 2018. 720–725.

Robertson Frankie. Word Discriminations for Vocabulary Inventory Prediction // Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). Held Online: INCOMA Ltd., IX 2021. 1188–1195.

Roman Roman Homero, Bisk Yonatan, Thomason Jesse, Celikyilmaz Asli, Gao Jianfeng. RMM: A Recursive Mental Model for Dialogue Navigation // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 1732–1745.

Rosenman Shachar, Jacovi Alon, Goldberg Yoav. Exposing Shallow Heuristics of Relation Extraction Models with Challenge Data // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 3702–3710.

Rossiello Gaetano, Gliozzo Alfio, Farrell Robert, Fauceglia Nicolas, Glass Michael. Learning Relational Representations by Analogy using Hierarchical Siamese Networks // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 3235–3245.

Rotman Guy, Feder Amir, Reichart Roi. Model Compression for Domain Adaptation through Causal Effect Estimation // Transactions of the Association for Computational Linguistics. 2021. 9. 1355–1373.

Roy Kalyani, Goyal Pawan, Pandey Manish. Attribute Value Generation from Product Title using Language Models // Proceedings of The 4th Workshop on e-Commerce and NLP. Online: Association for Computational Linguistics, VIII 2021. 13–17.

Roy Subhro, Roth Dan. Mapping to Declarative Knowledge for Word Problem Solving // Transactions of the Association for Computational Linguistics. 2018. 6. 159–172.

Rozen Ohad, Shwartz Vered, Aharoni Roee, Dagan Ido. Diversify Your Datasets: Analyzing Generalization via Controlled Variance in Adversarial Datasets // Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). Hong Kong, China: Association for Computational Linguistics, XI 2019. 196–205.

On learning the past tenses of English verbs. // . 1986.

Russin Jacob, Jo Jason, O'Reilly Randall, Bengio Yoshua. Compositional Generalization by Factorizing Alignment and Translation // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Online: Association for Computational Linguistics, VII 2020. 313–327.

Rybak Piotr, Mroczkowski Robert, Tracz Janusz, Gawlik Ireneusz. KLEJ: Comprehensive Benchmark for Polish Language Understanding // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 1191–1201.

Sachan Devendra, Zaheer Manzil, Salakhutdinov Ruslan. Investigating the Working of Text Classifiers // Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: Association for Computational Linguistics, VIII 2018. 2120–2131.

Saha Swarnadeep, Yadav Prateek, Bansal Mohit. multiPRover: Generating Multiple Proofs for Improved Interpretability in Rule Reasoning // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 3662–3677.

Salvatore Felipe, Finger Marcelo, Hirata Jr Roberto. A logical-based corpus for cross-lingual evaluation // Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019). Hong Kong, China: Association for Computational Linguistics, XI 2019. 22–30.

Sanchez Ivan, Mitchell Jeff, Riedel Sebastian. Behavior Analysis of NLI Models: Uncovering the Influence of Three Factors on Robustness // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, VI 2018. 1975–1985.

Sanh Victor, Webson Albert, Raffel Colin, Bach Stephen, Sutawika Lintang, Alyafeai Zaid, Chaffin Antoine, Stiegler Arnaud, Raja Arun, Dey Manan, Bari M Saiful, Xu Canwen, Thakker Urmish, Sharma Shanya Sharma, Szczechla Eliza, Kim Taewoon, Chhablani Gunjan, Nayak Nihal, Datta Debajyoti, Chang Jonathan, Jiang Mike Tian-Jian, Wang Han, Manica Matteo, Shen Sheng, Yong Zheng Xin, Pandey Harshit, Bawden Rachel, Wang Thomas, Neeraj Trishala, Rozen Jos, Sharma Abheesht, Santilli Andrea, Fevry Thibault, Fries Jason Alan, Teehan Ryan, Scao Teven Le, Biderman Stella, Gao Leo, Wolf Thomas, Rush Alexander M. Multitask Prompted Training Enables Zero-Shot Task Generalization // International Conference on Learning Representations. 2022.

Sauer Anna, Asaadi Shima, Küch Fabian. Knowledge Distillation Meets Few-Shot Learning: An Approach for Few-Shot Intent Classification Within and Across Domains // Proceedings of the 4th Workshop on NLP for Conversational AI. Dublin, Ireland: Association for Computational Linguistics, V 2022. 108–119.

Saxton David, Grefenstette Edward, Hill Felix, Kohli Pushmeet. Analysing Mathematical Reasoning Abilities of Neural Models // Proceedings of the 7th International Conference on Learning Representations (ICLR). 2019.

Scherbakov Andreas, Whittle Liam, Kumar Ritesh, Singh Siddharth, Coleman Matthew, Vylomova Ekaterina. Anlirika: An LSTM–CNN Flow Twister for Spoken Language Identification // Proceedings of the Third Workshop on Computational Typology and Multilingual NLP. Online: Association for Computational Linguistics, VI 2021. 145–148.

Schmidt Florian. Generalization in Generation: A closer look at Exposure Bias // Proceedings of the 3rd Workshop on Neural Generation and Translation. Hong Kong: Association for Computational Linguistics, XI 2019. 157–167.

Sen Indira, Samory Mattia, Flöck Fabian, Wagner Claudia, Augenstein Isabelle. How Does Counterfactually Augmented Data Impact Models for Social Computing Constructs? // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 325–344.

Sen Prithviraj, Li Yunyao, Kandogan Eser, Yang Yiwei, Lasecki Walter. HEIDL: Learning Linguistic Expressions with Deep Learning and Human-in-the-Loop // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Florence, Italy: Association for Computational Linguistics, VII 2019. 135–140.

Sengupta Ayan. DATAMAFIA at WNUT-2020 Task 2: A Study of Pre-trained Language Models along with Regularization Techniques for Downstream Tasks // Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). Online: Association for Computational Linguistics, XI 2020. 371–377.

Shaw Peter, Chang Ming-Wei, Pasupat Panupong, Toutanova Kristina. Compositional Generalization and Natural Language Variation: Can a Semantic Parsing Approach Handle Both? // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 922–938.

Shen Aili, Mistica Meladel, Salehi Bahar, Li Hang, Baldwin Timothy, Qi Jianzhong. Evaluating Document Coherence Modeling // Transactions of the Association for Computational Linguistics. 2021a. 9. 621–640.

Shen Yikang, Tan Shawn, Sordoni Alessandro, Reddy Siva, Courville Aaron. Explicitly Modeling Syntax in Language Models with Incremental Parsing and a Dynamic Oracle // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021b. 1660–1672.

Shen Yilin, Hsu Yen-Chang, Ray Avik, Jin Hongxia. Enhancing the generalization for Intent Classification and Out-of-Domain Detection in SLU // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021c. 2443–2453.

Shrivastava Akshat, Chuang Pierce, Babu Arun, Desai Shrey, Arora Abhinav, Zotov Alexander, Aly Ahmed. Span Pointer Networks for Non-Autoregressive Task-Oriented Semantic Parsing // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 1873–1886.

Shuster Kurt, Poff Spencer, Chen Moya, Kiela Douwe, Weston Jason. Retrieval Augmentation Reduces Hallucination in Conversation // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 3784–3803.

Shvartzshanider Yan, Balashankar Ananth, Wies Thomas, Subramanian Lakshminarayanan. RECIPE: Applying Open Domain Question Answering to Privacy Policies // Proceedings of the Workshop on Machine Reading for Question Answering. Melbourne, Australia: Association for Computational Linguistics, VII 2018. 71–77.

Shwartz Vered, Dagan Ido. Paraphrase to Explicate: Revealing Implicit Noun-Compound Relations // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, VII 2018. 1200–1211.

Silfverberg Miikka, Tyers Francis, Nicolai Garrett, Hulden Mans. Do RNN States Encode Abstract Phonological Alternations? // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 5501–5513.

Sinha Koustuv, Sodhani Shagun, Dong Jin, Pineau Joelle, Hamilton William L. CLUTRR: A Diagnostic Benchmark for Inductive Reasoning from Text // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 4506–4515.

Smith Shaden, Patwary Mostofa, Norick Brandon, LeGresley Patrick, Rajbhandari Samyam, Casper Jared, Liu Zhun, Prabhumoye Shrimai, Zerveas George, Korthikanti Vijay, Zhang Elton, Child Rewon, Aminabadi Reza Yazdani, Bernauer Julie, Song Xia, Shoeybi Mohammad, He Yuxiong, Houston Michael, Tiwary Saurabh, Catanzaro Bryan. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. 2022.

Søgaard Anders, Ebert Sebastian, Bastings Jasmijn, Filippova Katja. We Need To Talk About Random Splits // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, IV 2021. 1823–1832.

Srivastava Aarohi, Rastogi Abhinav, Rao Abhishek, Shoeb Abu Awal Md, Abid Abubakar, Fisch Adam, Brown Adam R, Santoro Adam, Gupta Aditya, Garriga-Alonso Adrià, others . Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models // CoRR. 2022. abs/2206.04615.

Srivastava Megha, Goodman Noah. Question Generation for Adaptive Education // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021. 692–701.

Stacey Joe, Minervini Pasquale, Dubossarsky Haim, Riedel Sebastian, Rocktäschel Tim. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 8281–8291.

Su Dan, Xu Yan, Winata Genta Indra, Xu Peng, Kim Hyeondey, Liu Zihan, Fung Pascale. Generalizing Question Answering System with Pre-trained Language Model Fine-tuning // Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Hong Kong, China: Association for Computational Linguistics, XI 2019. 203–211.

Su Ying, Zhang Hongming, Song Yangqiu, Zhang Tong. Rare and Zero-shot Word Sense Disambiguation using Z-Reweighting // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 4713–4723.

Subramanian Sanjay, Roth Dan. Improving Generalization in Coreference Resolution via Adversarial Training // Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 192–197.

Suhr Alane, Chang Ming-Wei, Shaw Peter, Lee Kenton. Exploring Unexplored Generalization Challenges for Cross-Database Semantic Parsing // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 8372–8388.

Sun Runxin, He Shizhu, Zhu Chong, He Yaohan, Li Jinlong, Zhao Jun, Liu Kang. Leveraging Explicit Lexico-logical Alignments in Text-to-SQL Parsing // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 283–289.

Sung Chul, Goel Vaibhava, Marcheret Etienne, Rennie Steven, Nahamoo David. CNNBiF: CNN-based Bigram Features for Named Entity Recognition // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 1016–1021.

Swayamdipta Swabha, Schwartz Roy, Lourie Nicholas, Wang Yizhong, Hajishirzi Hannaneh, Smith Noah A., Choi Yejin. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 9275–9293.

Takahashi Takumi, Taniguchi Motoki, Taniguchi Tomoki, Ohkuma Tomoko. CLER: Cross-task Learning with Expert Representation to Generalize Reading and Understanding // Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Hong Kong, China: Association for Computational Linguistics, XI 2019. 183–190.

Talat Zeerak, Thorne James, Bingel Joachim. Bridging the gaps: Multi task learning for domain transfer of hate speech detection // Online harassment. 2018. 29–55.

Talman Aarne, Chatzikyriakidis Stergios. Testing the Generalization Power of Neural Network Models across NLI Benchmarks // Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Florence, Italy: Association for Computational Linguistics, VIII 2019. 85–94.

Talmor Alon, Berant Jonathan. MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 4911–4921.

Talmor Alon, Elazar Yanai, Goldberg Yoav, Berant Jonathan. oLMpics-On What Language Model Pre-training Captures // Transactions of the Association for Computational Linguistics. 2020. 8. 743–758.

Tang Hongxuan, Li Hongyu, Liu Jing, Hong Yu, Wu Hua, Wang Haifeng. DuReader_robust: A Chinese Dataset Towards Evaluating Robustness and Generalization of Machine Reading Comprehension in Real-World Applications // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021. 955–963.

Tänzer Michael, Ruder Sebastian, Rei Marek. Memorisation versus Generalisation in Pre-trained Language Models // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 7564–7578.

Taya Yuki, Kanashiro Pereira Lis, Cheng Fei, Kobayashi Ichiro. OCHADAI-KYOTO at SemEval-2021 Task 1: Enhancing Model Generalization and Robustness for Lexical Complexity Prediction // Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). Online: Association for Computational Linguistics, VIII 2021. 17–23.

Tayyar Madabushi Harish, Kochkina Elena, Castelle Michael. Cost-Sensitive BERT for Generalisable Sentence Classification on Imbalanced Data // Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda. Hong Kong, China: Association for Computational Linguistics, XI 2019. 125–134.

Thorn Jakobsen Terne Sasha, Barrett Maria, Søgaard Anders. Spurious Correlations in Cross-Topic Argument Mining // Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics. Online: Association for Computational Linguistics, VIII 2021. 263–277.

Thrush Tristan, Wilcox Ethan, Levy Roger. Investigating Novel Verb Learning in BERT: Selectional Preference Classes and Alternation-Based Syntactic Generalization // Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Online: Association for Computational Linguistics, XI 2020. 265–275.

Tian Jidong, Li Yitian, Chen Wenqing, Xiao Liqiang, He Hao, Jin Yaohui. Diagnosing the First-Order Logical Reasoning Ability Through LogicNLI // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 3738–3747.

Toshniwal Shubham, Xia Patrick, Wiseman Sam, Livescu Karen, Gimpel Kevin. On Generalization in Coreference Resolution // Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 111–120.

Tran Van-Khanh, Nguyen Le-Minh. Natural Language Generation for Spoken Dialogue System using RNN Encoder-Decoder Networks // Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada: Association for Computational Linguistics, VIII 2017. 442–451.

Tu Lifu, Lalwani Garima, Gella Spandana, He He. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models // Transactions of the Association for Computational Linguistics. 2020. 8. 621–633.

Vilar David, Federico Marcello. A Statistical Extension of Byte-Pair Encoding // Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021). Bangkok, Thailand (online): Association for Computational Linguistics, VIII 2021. 263–275.

Vu Thuy-Trang, Khadivi Shahram, Phung Dinh, Haffari Gholamreza. Domain Generalisation of NMT: Fusing Adapters with Leave-One-Domain-Out Training // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 582–588.

Vulić Ivan, Ponzetto Simone Paolo, Glavaš Goran. Multilingual and Cross-Lingual Graded Lexical Entailment // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 4963–4974.

Vylomova Ekaterina, White Jennifer, Salesky Elizabeth, Mielke Sabrina J., Wu Shijie, Ponti Edoardo Maria, Hall Maudslay Rowan, Zmigrod Ran, Valvoda Josef, Toldova Svetlana, Tyers Francis, Klyachko Elena, Yegorov Ilya, Krizhanovsky Natalia, Czarnowska Paula, Nikkarinen Irene, Krizhanovsky Andrew, Pimentel Tiago, Torroba Hennigen Lucas, Kirov Christo, Nicolai Garrett, Williams Adina, Anastasopoulos Antonios, Cruz Hilaria, Chodroff Eleanor, Cotterell Ryan, Silfverberg Miikka, Hulden Mans. SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection // Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. Online: Association for Computational Linguistics, VII 2020. 1–39.

Wadhwa Soumya, Embar Varsha, Grabmair Matthias, Nyberg Eric. Towards Inference-Oriented Reading Comprehension: ParallelQA // Proceedings of the Workshop on Generalization in the Age of Deep Learning. New Orleans, Louisiana: Association for Computational Linguistics, VI 2018. 1–7.

Wang Alex, Pruksachatkun Yada, Nangia Nikita, Singh Amanpreet, Michael Julian, Hill Felix, Levy Omer, Bowman Samuel. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems // Advances in Neural Information Processing Systems. 32. 2019a.

Wang Bailin, Lapata Mirella, Titov Ivan. Meta-Learning for Domain Generalization in Semantic Parsing // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021a. 366–379.

Wang Bailin, Shin Richard, Liu Xiaodong, Polozov Oleksandr, Richardson Matthew. RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020a. 7567–7578.

Wang Bailin, Titov Ivan, Lapata Mirella. Learning Semantic Parsers from Denotations with Latent Structured Alignments and Abstract Programs // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019b. 3774–3785.

Wang Bailin, Yin Wenpeng, Lin Xi Victoria, Xiong Caiming. Learning to Synthesize Data for Semantic Parsing // Proceedings of the 2021 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021b. 2760–2766.

Wang Chunliu, Noord Rik van, Bisazza Arianna, Bos Johan. Evaluating Text Generation from Discourse Representation Structures // Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021). Online: Association for Computational Linguistics, VIII 2021c. 73–83.

Wang Yizhong, Mishra Swaroop, Alipoormolabashi Pegah, Kordi Yeganeh, Mirzaei Amirreza, Arunkumar Anjana, Ashok Arjun, Dhanasekaran Arut Selvan, Naik Atharva, Stap David, Pathak Eshaan, Karamanolakis Giannis, Lai Haizhi Gary, Purohit Ishan, Mondal Ishani, Anderson Jacob, Kuznia Kirby, Doshi Krima, Patel Maitreya, Pal Kuntal Kumar, Moradshahi Mehrad, Parmar Mihir, Purohit Mirali, Varshney Neeraj, Kaza Phani Rohitha, Verma Pulkit, Puri Ravsehaj Singh, Karia Rushang, Sampat Shailaja Keyur, Doshi Savan, Mishra Siddhartha, Reddy Sujan, Patro Sumanta, Dixit Tanay, Shen Xudong, Baral Chitta, Choi Yejin, Smith Noah A., Hajishirzi Hannaneh, Khashabi Daniel. Benchmarking Generalization via In-Context Instructions on 1,600+ Language Tasks. 2022.

Wang Zirui, Lipton Zachary C., Tsvetkov Yulia. On Negative Interference in Multilingual Models: Findings and A Meta-Learning Treatment // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020b. 4438–4450.

Warstadt Alex, Singh Amanpreet, Bowman Samuel R. Neural Network Acceptability Judgments // Transactions of the Association for Computational Linguistics. 2019. 7. 625–641.

Warstadt Alex, Zhang Yian, Li Xiaocheng, Liu Haokun, Bowman Samuel R. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually) // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 217–235.

Weber Leon, Münchmeyer Jannes, Garda Samuele, Leser Ulf. Extend, don't rebuild: Phrasing conditional graph modification as autoregressive sequence labelling // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021a. 1213–1224.

Weber Lucas, Jumelet Jaap, Bruni Elia, Hupkes Dieuwke. Language Modelling as a Multi-Task Problem // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, IV 2021b. 2049–2060.

Weber Noah, Shekhar Leena, Balasubramanian Niranjan. The Fine Line between Linguistic Generalization and Failure in Seq2Seq-Attention Models // Proceedings of the Workshop on Generalization in the Age of Deep Learning. New Orleans, Louisiana: Association for Computational Linguistics, VI 2018. 24–27.

Wei Jason, Bosma Maarten, Zhao Vincent, Guu Kelvin, Yu Adams Wei, Lester Brian, Du Nan, Dai Andrew M., Le Quoc V. Finetuned Language Models are Zero-Shot Learners // International Conference on Learning Representations. 2022.

Wei Jason, Garrette Dan, Linzen Tal, Pavlick Ellie. Frequency Effects on Syntactic Rule Learning in Transformers // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 932–948.

Welbl Johannes, Minervini Pasquale, Bartolo Max, Stenetorp Pontus, Riedel Sebastian. Undersensitivity in Neural Reading Comprehension // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 1152–1165.

Weller Orion, Lourie Nicholas, Gardner Matt, Peters Matthew E. Learning from Task Descriptions // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 1361–1375.

Wilcox Ethan, Qian Peng, Futrell Richard, Ballesteros Miguel, Levy Roger. Structural Supervision Improves Learning of Non-Local Grammatical Dependencies // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019. 3302–3312.

Wilcox Ethan, Qian Peng, Futrell Richard, Kohita Ryosuke, Levy Roger, Ballesteros Miguel. Structural Supervision Improves Few-Shot Learning and Syntactic Generalization in Neural Language Models // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 4640–4652.

Winata Genta Indra, Lin Zhaojiang, Fung Pascale. Learning Multilingual Meta-Embeddings for Code-Switching Named Entity Recognition // Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). Florence, Italy: Association for Computational Linguistics, VIII 2019. 181–186.

Wong Francis CK, Wang William SY. Generalisation towards combinatorial productivity in language acquisition by simple recurrent networks // 2007 International Conference on Integration of Knowledge Intensive Multi-Agent Systems. 2007. 139–144.

Wu Bowen, Huang Haoyang, Wang Zongsheng, Feng Qihang, Yu Jingsong, Wang Baoxun. Improving the Robustness of Deep Reading Comprehension Models by Leveraging Syntax Prior // Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Hong Kong, China: Association for Computational Linguistics, XI 2019. 53–57.

Wu Mingzhu, Moosavi Nafise Sadat, Rücklé Andreas, Gurevych Iryna. Improving QA Generalization by Concurrent Modeling of Multiple Biases // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 839–853.

Wu Shijie, Dredze Mark. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 833–844.

Wu Yuan, Inkpen Diana, El-Roby Ahmed. Conditional Adversarial Networks for Multi-Domain Text Classification // Proceedings of the Second Workshop on Domain Adaptation for NLP. Kyiv, Ukraine: Association for Computational Linguistics, IV 2021a. 16–27.

Wu Zeqiu, Lu Bo-Ru, Hajishirzi Hannaneh, Ostendorf Mari. DIALKI: Knowledge Identification in Conversational Systems through Dialogue-Document Contextualization // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021b. 1852–1863.

Wullach Tomer, Adler Amir, Minkov Einat. Fight Fire with Fire: Fine-tuning Hate Detectors using Large Samples of Generated Hate Speech // Findings of the Association for Computational Linguistics:

EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 4699–4705.

Xia Congying, Xiong Caiming, Yu Philip, Socher Richard. Composed Variational Natural Language Generation for Few-shot Intents // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 3379–3388.

Xiachong Feng, Xiaocheng Feng, Bing Qin. Incorporating Commonsense Knowledge into Abstractive Dialogue Summarization via Heterogeneous Graph Networks // Proceedings of the 20th Chinese National Conference on Computational Linguistics. Huhhot, China: Chinese Information Processing Society of China, VIII 2021. 964–975.

Xiao Liqiang, Wang Lu, He Hao, Jin Yaohui. Modeling Content Importance for Summarization with Pre-trained Language Models // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020. 3606–3611.

Xie Tianbao, Wu Chen Henry, Shi Peng, Zhong Ruiqi, Scholak Torsten, Yasunaga Michihiro, Wu Chien-Sheng, Zhong Ming, Yin Pengcheng, Wang Sida I., Zhong Victor, Wang Bailin, Li Chengzu, Boyle Connor, Ni Ansong, Yao Ziyu, Radev Dragomir, Xiong Caiming, Kong Lingpeng, Zhang Rui, Smith Noah A., Zettlemoyer Luke, Yu Tao. UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models // CoRR. 2022. abs/2201.05966.

Xin Ji, Xiong Chenyan, Srinivasan Ashwin, Sharma Ankita, Jose Damien, Bennett Paul. Zero-Shot Dense Retrieval with Momentum Adversarial Domain Invariant Representations // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 4008–4020.

Xu Chang, Paris Cécile, Nepal Surya, Sparks Ross. Cross-Target Stance Classification with Self-Attention Networks // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics, VII 2018. 778–783.

Xu Hanwei, Chen Yujun, Du Yulun, Shao Nan, Wang Yanggang, Li Haiyu, Yang Zhilin. ZeroPrompt: Scaling Prompt-Based Pretraining to 1, 000 Tasks Improves Zero-Shot Generalization // CoRR. 2022. abs/2201.06910.

Xu Peng, Kumar Dhruv, Yang Wei, Zi Wenjie, Tang Keyi, Huang Chenyang, Cheung Jackie Chi Kit, Prince Simon J.D., Cao Yanshuai. Optimizing Deeper Transformers on Small Datasets // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021a. 2089–2102.

Xu Peng, Saghir Hamidreza, Kang Jin Sung, Long Teng, Bose Avishek Joey, Cao Yanshuai, Cheung Jackie Chi Kit. A Cross-Domain Transferable Neural Coherence Model // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 678–687.

Xu Runxin, Luo Fuli, Zhang Zhiyuan, Tan Chuanqi, Chang Baobao, Huang Songfang, Huang Fei. Raise a Child in Large Language Model: Towards Effective and Generalizable Fine-tuning // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021b. 9514–9528.

Yaghoobzadeh Yadollah, Mehri Soroush, Combes Remi Tachet des, Hazen T. J., Sordoni Alessandro. Increasing Robustness to Spurious Correlations using Forgettable Examples // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, IV 2021. 3319–3332.

Yanaka Hitomi, Mineshima Koji. Assessing the Generalization Capacity of Pre-trained Language Models through Japanese Adversarial Natural Language Inference // Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 337–349.

Yanaka Hitomi, Mineshima Koji, Bekki Daisuke, Inui Kentaro. Do Neural Models Learn Systematicity of Monotonicity Inference in Natural Language? // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 6105–6117.

Yanaka Hitomi, Mineshima Koji, Bekki Daisuke, Inui Kentaro, Sekine Satoshi, Abzianidze Lasha, Bos Johan. Can Neural Networks Understand Monotonicity Reasoning? // Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Florence, Italy: Association for Computational Linguistics, VIII 2019. 31–40.

Yanaka Hitomi, Mineshima Koji, Inui Kentaro. Exploring Transitivity in Neural NLI Models through Veridicality // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, IV 2021. 920–934.

Yang Fan, Mukherjee Arjun, Zhang Yifan. Leveraging Multiple Domains for Sentiment Classification // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan: The COLING 2016 Organizing Committee, XII 2016. 2978–2988.

Yang Kaiyu, Deng Jia. Strongly incremental constituency parsing with graph neural networks // Advances in Neural Information Processing Systems. 2020. 33. 21687–21698.

Yang Sen, Cui Leyang, Ning Ruoxi, Wu Di, Zhang Yue. Challenges to Open-Domain Constituency Parsing // Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, V 2022. 112–127.

Yang Yiben, Malaviya Chaitanya, Fernandez Jared, Swayamdipta Swabha, Le Bras Ronan, Wang Ji-Ping, Bhagavatula Chandra, Choi Yejin, Downey Doug. Generative Data Augmentation for Commonsense Reasoning // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 1008–1025.

Ye Qinyuan, Huang Xiao, Boschee Elizabeth, Ren Xiang. Teaching Machine Comprehension with Compositional Explanations // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020. 1599–1615.

Ye Qinyuan, Lin Bill Yuchen, Ren Xiang. CrossFit: A Few-shot Learning Challenge for Cross-task Generalization in NLP // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 7163–7189.

Ye Qinyuan, Ren Xiang. Learning to Generate Task-Specific Adapters from Task Description // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online: Association for Computational Linguistics, VIII 2021. 646–653.

Ye Xi, Yavuz Semih, Hashimoto Kazuma, Zhou Yingbo, Xiong Caiming. RNG-KBQA: Generation Augmented Iterative Ranking for Knowledge Base Question Answering // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 6032–6043.

Yin Pengcheng, Fang Hao, Neubig Graham, Pauls Adam, Platanios Emmanouil Antonios, Su Yu, Thomson Sam, Andreas Jacob. Compositional Generalization for Neural Semantic Parsing via Span-level Supervised Attention // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 2810–2823.

Yin Wenpeng, Rajani Nazneen Fatema, Radev Dragomir, Socher Richard, Xiong Caiming. Universal Natural Language Processing with Limited Annotations: Try Few-shot Textual Entailment as a Start // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, XI 2020a. 8229–8239.

Yin Wenpeng, Roth Dan, Schütze Hinrich. End-Task Oriented Textual Entailment via Deep Explorations of Inter-Sentence Interactions // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia: Association for Computational Linguistics, VII 2018. 540–545.

Yin Xusen, Weischedel Ralph, May Jonathan. Learning to Generalize for Sequential Decision Making // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020b. 3046–3063.

Yogatama Dani, d'Autume Cyprien de Masson, Connor Jerome, Kocisky Tomas, Chrzanowski Mike, Kong Lingpeng, Lazaridou Angeliki, Ling Wang, Yu Lei, Dyer Chris, others . Learning and evaluating general linguistic intelligence // arXiv preprint arXiv:1901.11373. 2019.

Yoo Jin Yong, Qi Yanjun. Towards Improving Adversarial Training of NLP Models // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 945–956.

Yu Tao, Joty Shafiq. Effective Fine-Tuning Methods for Cross-lingual Adaptation // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 8492–8501.

Yu Tao, Zhang Rui, Yang Kai, Yasunaga Michihiro, Wang Dongxu, Li Zifan, Ma James, Li Irene, Yao Qingning, Roman Shanelle, Zhang Zilin, Radev Dragomir. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, X-XI 2018. 3911–3921.

Yu Tao, Zhang Rui, Yasunaga Michihiro, Tan Yi Chern, Lin Xi Victoria, Li Suyi, Er Heyang, Li Irene, Pang Bo, Chen Tao, Ji Emily, Dixit Shreya, Proctor David, Shim Sungrok, Kraft Jonathan, Zhang Vincent, Xiong Caiming, Socher Richard, Radev Dragomir. SParC: Cross-Domain Semantic Parsing in Context // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019a. 4511–4523.

Yu Tiezheng, Liu Zihan, Fung Pascale. AdaptSum: Towards Low-Resource Domain Adaptation for Abstractive Summarization // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021. 5892–5904.

Yu Yue, Zhu Yilun, Liu Yang, Liu Yan, Peng Siyao, Gong Mackenzie, Zeldes Amir. GumDrop at the DISRPT2019 Shared Task: A Model Stacking Approach to Discourse Unit Segmentation and Connective Detection // Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019. Minneapolis, MN: Association for Computational Linguistics, VI 2019b. 133–143.

Yue Xiang, Jimenez Gutierrez Bernal, Sun Huan. Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 4474–4486.

Zaharia George-Eduard, Smădu Răzvan-Alexandru, Cercel Dumitru, Dascalu Mihai. Domain Adaptation in Multilingual and Multi-Domain Monolingual Settings for Complex Word Identification // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 70–80.

Zeng Yutao, Jin Xiaolong, Guan Saiping, Guo Jiafeng, Cheng Xueqi. Event Coreference Resolution with their Paraphrases and Argument-aware Embeddings // Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online): International Committee on Computational Linguistics, XII 2020. 3084–3094.

Zhang Dejiao, Nallapati Ramesh, Zhu Henghui, Nan Feng, Santos Cicero Nogueira dos, McKeown Kathleen, Xiang Bing. Margin-aware Unsupervised Domain Adaptation for Cross-lingual Text Labeling // Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, XI 2020a. 3527–3536.

Zhang Haode, Zhang Yuwei, Zhan Li-Ming, Chen Jiaxin, Shi Guangyuan, Wu Xiao-Ming, Lam Albert Y.S. Effectiveness of Pre-training for Few-shot Intent Classification // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021a. 1114–1120.

Zhang Hongming, Song Yan, Song Yangqiu, Yu Dong. Knowledge-aware Pronoun Coreference Resolution // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 867–876.

Zhang Michael, Choi Eunsol. SituatedQA: Incorporating Extra-Linguistic Contexts into QA // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 7371–7387.

Zhang Mozhi, Fujinuma Yoshinari, Paul Michael J., Boyd-Graber Jordan. Why Overfitting Isn't Always Bad: Retrofitting Cross-Lingual Word Embeddings to Dictionaries // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020b. 2214–2220.

Zhang Mozhi, Wang Wei, Deb Budhaditya, Zheng Guoqing, Shokouhi Milad, Awadallah Ahmed Hassan. A Dataset and Baselines for Multilingual Reply Suggestion // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021b. 1207–1220.

Zhang Susan, Roller Stephen, Goyal Naman, Artetxe Mikel, Chen Moya, Chen Shuohui, Dewan Christopher, Diab Mona, Li Xian, Lin Xi Victoria, Mihaylov Todor, Ott Myle, Shleifer Sam, Shuster Kurt, Simig Daniel, Koura Punit Singh, Sridhar Anjali, Wang Tianlu, Zettlemoyer Luke. OPT: Open Pre-trained Transformer Language Models. 2022a.

Zhang Yiwen, Yuan Caixia, Wang Xiaojie, Bai Ziwei, Liu Yongbin. Learn to Adapt for Generalized Zero-Shot Text Classification // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022b. 517–527.

Zhao Chen, Su Yu, Pauls Adam, Platanios Emmanouil Antonios. Bridging the Generalization Gap in Text-to-SQL Parsing with Schema Expansion // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 5568–5578.

Zhao Tiancheng, Eskenazi Maxine. Zero-Shot Dialog Generation with Cross-Domain Latent Actions // Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue. Melbourne, Australia: Association for Computational Linguistics, VII 2018. 1–10.

Zhao Tiancheng, Lu Xiaopeng, Lee Kyusong. SPARTA: Efficient Open-Domain Question Answering via Sparse Transformer Matching Retrieval // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021a. 565–575.

Zhao Tianyu, Lala Divesh, Kawahara Tatsuya. Designing Precise and Robust Dialogue Response Evaluators // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, VII 2020. 26–33.

Zhao Wei, Peyrard Maxime, Liu Fei, Gao Yang, Meyer Christian M., Eger Steffen. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, XI 2019. 563–578.

Zhao Xinyu, Lin Shih-Ting, Durrett Greg. Effective Distant Supervision for Temporal Relation Extraction // Proceedings of the Second Workshop on Domain Adaptation for NLP. Kyiv, Ukraine: Association for Computational Linguistics, IV 2021b. 195–203.

Zhao Yingzhu, Ni Chongjia, Leung Cheung-Chi, Joty Shafiq, Chng Eng Siong, Ma Bin. A Unified Speaker Adaptation Approach for ASR // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021c. 9339–9349.

Zheng Hao, Lapata Mirella. Compositional Generalization via Semantic Tagging // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 1022–1032.

Zheng Hao, Lapata Mirella. Disentangled Sequence to Sequence Learning for Compositional Generalization // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 4256–4268.

Zheng Shun, Han Xu, Lin Yankai, Yu Peilin, Chen Lu, Huang Ling, Liu Zhiyuan, Xu Wei. DIAG-NRE: A Neural Pattern Diagnosis Framework for Distantly Supervised Neural Relation Extraction // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019. 1419–1429.

Zhong Yang, Yang Jingfeng, Xu Wei, Yang Diyi. WIKIBIAS: Detecting Multi-Span Subjective Biases in Language // Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 1799–1814.

Zhou Ben, Khashabi Daniel, Tsai Chen-Tse, Roth Dan. Zero-Shot Open Entity Typing as Type-Compatible Grounding // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, X-XI 2018. 2065–2076.

Zhou Ben, Richardson Kyle, Ning Qiang, Khot Tushar, Sabharwal Ashish, Roth Dan. Temporal Reasoning on Implicit Events from Distant Supervision // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, VI 2021a. 1361–1371.

Zhou Chunting, Ma Xuezhe, Wang Di, Neubig Graham. Density Matching for Bilingual Word Embedding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, VI 2019a. 1588–1598.

Zhou Giulio, Lampouras Gerasimos. Generalising Multilingual Concept-to-Text NLG with Language Agnostic Delexicalisation // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, VIII 2021. 114–127.

Zhou Joey Tianyi, Zhang Hao, Jin Di, Zhu Hongyuan, Fang Meng, Goh Rick Siow Mong, Kwok Kenneth. Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, VII 2019b. 3461–3471.

Zhou Pei, Khanna Rahul, Lee Seyeon, Lin Bill Yuchen, Ho Daniel, Pujara Jay, Ren Xiang. RICA: Evaluating Robust Inference Capabilities Based on Commonsense Axioms // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021b. 7560–7579.

Zhou Xiang, Elfardy Heba, Christodoulopoulos Christos, Butler Thomas, Bansal Mohit. Hidden Biases in Unreliable News Detection Datasets // Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, IV 2021c. 2482–2492.

Zhou Yangqiaoyu, Tan Chenhao. Investigating the Effect of Natural Language Explanations on Out-of-Distribution Generalization in Few-shot NLI // Proceedings of the Second Workshop on Insights from Negative Results in NLP. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 117–124.

Zhou Yucheng, Shen Tao, Geng Xiubo, Long Guodong, Jiang Daxin. ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, V 2022. 2559–2575.

Zhou Zili, Valentino Marco, Landers Donal, Freitas André. Encoding Explanatory Knowledge for Zero-shot Science Question Answering // Proceedings of the 14th International Conference on Computational Semantics (IWCS). Groningen, The Netherlands (online): Association for Computational Linguistics, VI 2021d. 38–50.

Ziser Yftah, Reichart Roi. Neural Structural Correspondence Learning for Domain Adaptation // Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada: Association for Computational Linguistics, VIII 2017. 400–410.

Zou Yicheng, Zhu Bolin, Hu Xingwu, Gui Tao, Zhang Qi. Low-Resource Dialogue Summarization with Domain-Agnostic Multi-Source Pretraining // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 80–91.

Zuo Simiao, Liang Chen, Jiang Haoming, Liu Xiaodong, He Pengcheng, Gao Jianfeng, Chen Weizhu, Zhao Tuo. Adversarial Regularization as Stackelberg Game: An Unrolled Optimization Approach // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, XI 2021. 6562–6577.