Towards Domain Robustness of Neural Language Models

Michal Štefánik and Petr Sojka

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
stefanik.m@mail.muni.cz, sojka@fi.muni.cz

Abstract. This work summarises recent progress in generalization evaluation and training of deep neural networks, categorized in data-centric and model-centric overviews. Grounded in the results of the referenced work, we propose three future directions towards reaching higher robustness of language models to an unknown domain or its adaptation to an existing domain of interest. In the example propositions that practically complement each of the directions, we introduce novel ideas of **a**) dynamic objective selection, **b**) language modeling respecting the token similarities to the ground truth and **c**) a framework of additive component of the loss utilizing the well-performing generalization measures.

Keywords: Generalization · Debiasing · Domain extrapolation · Domain adaptation · Domain robustness · Neural language models

> "Education is the most powerful weapon we can use to change the world."
> Nelson Mandela

1 Introduction

Deep language models have found their application in a wide variety of tasks, ranging among other aspects in their semantic complexity and a domain of applicability. While a domain of some applications can be bound, commonly, we can not afford to utilize a specialized model for every possible *domain*, i.e., a set of samples of which we apply the language model, conditioned by a distinct situational and pragmatic background. Furthermore, our domains of interest might not even be preliminary known, as is often the native case in generative tasks, such as neural machine translation, summarization, or paraphrasing; think, for example, of a variety of domains for which a general-purpose machine translation system can be applied.

The exceeded reliance of the models on characteristics of a single training domain shows as an increasing problem only with an increased expressivity of the deep architectures, which are for the first time able to accurately model the non-representative relations not easily apparent to their maintainer. As one of the first, McCoy [24] demonstrates a reliance of state-of-the-art transformer

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, pp. 91–103, 2021. © Tribun EU 2021

model on heuristical shortcuts on language inference [42], specifically on a lexical and subsequence overlap between the premise and hypothesis. Belinkov [2] and Berard [4] show fragility of neural machine translation models to typos and misspelling, and vocabulary shift, respectively, both common for non-canonical domains that the systems are usually not trained on. A large branch of work follows, either in aims to empirically identify domain-specific biases in commonly-used data sets [39,14,29,16], or in aims to heuristically eliminate these biases in data [24,27,48].

This paper brings an introductory overview of the limited set of existing methods that address the qualitative discrepancy of applying the model to samples of different domain(s), regardless of the specific type of domain shift between the training and target domain.

Section 2, overviews the existing methods based on resampling the training domain samples or exposing the domain shift by using the data from two different domains. Further, in Section 3, we extend this list for a domain that *adjust* the standard training process via adjusting the objective of the training process.

Finally, in Section 4 we outline the open ends implied by the results of the preceding studies, which could lead to an enhancement of the model's domain robustness. We aim to describe these common directions tangibly enough to be utilizable in future research. We thoroughly describe a single technical proposition for each of the three outlined directions and leave its empirical evaluation to the subsequent studies.

2 Extrapolation using Data

Data approaches aim to utilize the available samples, possibly categorized by their domain of origin, in order to minimize test loss on samples of the domain of interest. In the scope of a well-recognized branch of work labeled as *domain adaptation*, the training situation is denoted by the availability of source domain X_s, which can be interpreted as a random variable generating the samples x_s with their corresponding labels y_s. Further we denote a target domain X_t, i.e. a domain of application, with a limited amount of $(x_t, y_t) \in X_t$, where it holds that $|X_t| < |X_s|$, or in some situations, where the amount of $y_t \in X_t$ is limited.

In the more extreme case referred in the literature as an evaluation for *domain generalization*, we restrict the training process to access only samples of source domain(s) X_s, and the samples of X_t are a priori unavailable. Arguably, this situation better corresponds to open-domain applications such as open machine translation.

2.1 Impact of Data Subsampling

Data Selection approaches aim to resample the samples used for the training process in order to maximize the generalization ability of the eventual model.

Denoising strategies elaborate on a hypothesis that some samples are less representative to the task of interest than others. Among more straightforward approaches, Lin [22] picks the "clean" set of samples according to their perplexity to the linear base model, keeping in the training set only the ones with low perplexity. Later, Moore [25] seeks to pick the samples x_s that minimally affect the sum log-likelihood of the model updated according to x_s. Similarly, Yarowsky [45] pick the training subsample based on a threshold on the sample output *confidence*. Zhou [49] iteratively applies the same strategy using an ensemble of three estimators, only picking the top-n most-confident samples, possibly avoiding the mangle confidence calibration, and refers to this approach to as *tri-training*.

An interesting, yet more complex approach, referred to as *Product-of-Experts* is introduced by [15]. Here, an ensemble of relatively small classifiers is used to *debias* the training samples by computing a dot product of class-wise *logits* of the ensemble and possibly discarding the samples for which the ensemble disagrees the most. Sanh [33] applies this approach to the training transformers model and finds interesting performance gains on out-of-domain performance. Similarly, Utama [39] identify the possibly-biased samples as the ones reaching high confidence only for a single one of the ensembled models and consecutively *weights* the training samples by their chance of exposing bias. In the broader scope, these approaches fit well into the PAC-Bayesian framework [40], roughly stating that if for the selected model M empirical error bound ϵ_M, then for the error for an *ensemble* E of such models it holds that $\epsilon_E \leq \epsilon_M$.

2.2 Ability to Distinguish Domains

Another approach to domain generalization leads through an *exposition* of the domain discrepancies, which is a necessary precondition for the model to comprehend and possibly to model it. This is theoretically supported by the work of Locatello [23], concluding that *distributional robustness* is not possible without the exposition of both *data* and *model* inductive biases. Bengio [3] demonstrates how these biases can be utilized by the model to fit the causal structure of the data and evaluate this ability in the situation where the data-specific inductive biases are known.

There are simpler ways how domain discrepancies can be effectively communicated to the model. For example, Shah [34] minimizes the Wasserstein distance of internal model representations between the samples of source and target domain, X_s and X_t. Jiang [17] first trains the domain classifier C_d distinguishing domains X_s and X_t and subsequently *weights* the samples $x_s \in X_s$ in the training by their correspondence to X_t as given by the confidence of C_d. Chadha [5] enhances out-of-domain performance of adapted model by adding so-called *maximum mean discrepancy loss* to the training objective, given by $\max(dist(x_s, x_t)) : x_s \in X_s, x_t \in X_t$.

3 Extrapolation and Training Process

The adjustments to the training process have proved to increase the distributional robustness of the final model in different variations. We identify that the authors of empirically-successful works in generalization use the regularization element, which corresponds to a specific well-performing generalization measure. Hence, we first describe popular evaluation measures and then describe the specific adjustments of the training process leading to a model with better generalization.

3.1 Evaluation of Extrapolation

In a large-scale study on image classification, Jiang [18] shows that the measures of so-called *spectral graph complexity* [28], *sharpness* of the parametrized space [19], or PAC-Bayesian measures [40], similar to the introduced Product-of-Experts, correlate the highest to the empirical out-of-domain performance of the convolutional model. Later, Dziugaite [11] dispute some of these results, reproducing the experiments in enhanced, fine-grained methodology, showing that the high average correlations of some measures, such as the spectral complexity, systematically fail under specific domain shifts.

Perhaps surprisingly, these studies agree upon the low correlations of the standard regularization techniques such as dropout or norms regularization, suggesting that an application of techniques sufficient to avoid in-domain overfitting might not be sufficient for reaching distributional robustness.

3.2 Training Process Adjustments

A large branch of studies shows that regularizing the training process using the referenced generalization measures positively impacts the distributional robustness of the model. However, note that most of the following studies were applied in evaluating image classification, with questionable relevance to transfer learning settings.

Barlett [1] uses spectral complexity as a norm in the training process of the AlexNet convolutional network and theoretically demonstrates that this property corresponds to the network generalization ability. Similarly, Foret [12] uses sharpness as an additive term of loss, computed on locally-surrounding inputs as an additive component of the training loss. In addition to increasing out-of-domain accuracy, the resulting model demonstrates higher robustness to noisy training in-domain samples. Referring to the process as "debiasing", Utama [39] utilize the commonly-evaluated PAC Bayesian confidence estimate in predictions in loss weighting.

Other adjustments give some insights into the impact of the composition of transfer learning objectives. While Teney [37] or Wang [26] demonstrate the cases where adaptation to a single domain harms out-of-distribution robustness of the model, Wu [43] concludes that adapting to multiple data sets can enhance the end model generalization. Additionally, Tu [38] reporting a positive impact of multitask learning to model's out-of-distribution accuracy, or by Xie [44] for additive consistency regularization in the training objective.

4 Future Perspectives

Grounded in the referenced studies and results, we now describe three potential directions that could mitigate the exposition of inductive biases in the language models and, consequently, reach their higher generalization ability. We enrich each one of these directions with a practical proposition that contributes to the described direction.

Overall, we observe that the strategy of interaction with a model during the training has a significant impact on the model's generalization ability, just like the teacher's methods and interaction have a principal effect on the student's performance. All of the introduced directions elaborate on interaction strategies towards the model on training time.

4.1 Impact of Objectives Curricula

"If we examine ourselves, we see that our faculties grow in such a manner that what goes before paves the way for what comes after." J. A. Comenius [8]

While many of the mentioned studies, for example, [5,43,38] enrich the training objective with an exposition of the domain discrepancies and their respective biases with reported positive impact to generalization, it is not clear how the specific strategies of doing so vary in effectiveness and efficiency. For instance, Gururangan [13] concludes that it is always beneficial to perform a fine-tuning to a domain or a task of interest by sequentially applying the different objectives, Tu [38] apply a concurrent objective schedule. Additionally, as some objectives might be easier than others, it is likely that some objectives overweight others over time, mitigating the further convergence, possibly necessary for learning the corner cases [38].

We propose to systematically enhance our comprehension of the performance of models in the different objectives: do we somewhat loose grasp of a general language understanding, reflected, for example, in Masked or Causal language modeling accuracy [10,31], or Denoising [21], when fine-tuning for a token or sequence classification on end task? If this degradation is significant, as suggested, for example, by the results of Popel [30], it motivates the results for a more complicated schedule of an application of objectives.

> *If a fine-tuning on end objective degrade performance of other relevant objectives, we are motivated to utilize a non-sequential schedule of these objectives in the common adaptations.*

We propose to confront a standard sequential schedule of the optimization of the objective with the novel ones. We aim to investigate at least the two strategies outlined in Figure 1: a "striped" schedule strategy, where the loss of *all* objectives

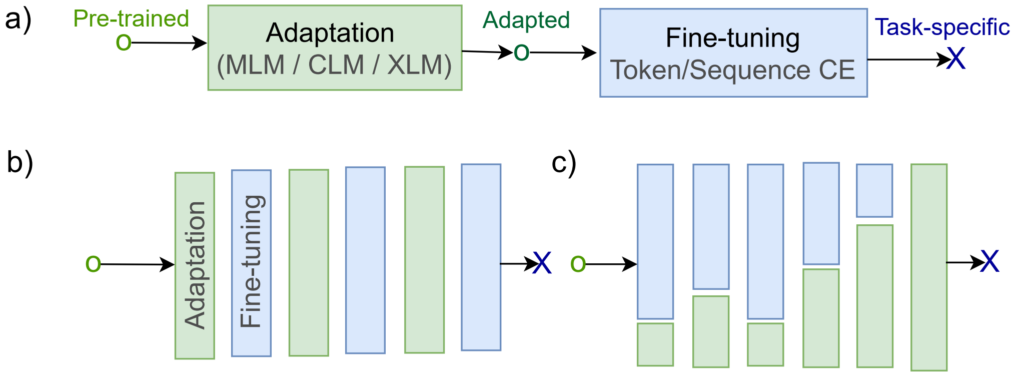

Fig. 1: Illustrative comparison of basic objective sampling strategies. Traditionally, domain adaptation is performed in sequential strategy (**a**). Presumably, a combined sampling strategy (**b**), could avoid performance decay of the unscheduled, yet relevant objective(s), as reported for instance by Popel [30]. A dynamic sampling (**c**), based for example on a state of the validation loss, could further eliminate this performance decay.

is included in each training step, and a candidate of the groups of "dynamic" strategies, where the objective selection is determined by a heuristic based on the immediate loss of given objective.

4.2 Softer Objectives

> "The proper education of the young does not consist in stuffing their heads with a mass of words, sentences, and ideas dragged together out of various authors, but in opening up their understanding to the outer world, so that a living stream may flow from their own minds, just as leaves, flowers, and fruit spring from the bud on a tree." J. A. Comenius [8]

The continuous over-parametrization of deep language models brings qualitative gains even by following the same, well-established objectives on the same, limited amount of training resources of end tasks, as shown for instance by [10,9]. Still, it makes sense to ask whether the commonly-used objectives expose the characteristics of the learned task in an *efficient* manner, both with respect to the computational resources and often expensive supervised data resources.

Consider the cases of Masked, or Causal language modeling, where 15% of randomly-selected tokens is masked. Presuming the Zipf law holding for the natural language artifacts in all its levels (from morphology to semantic of, e.g., coreference or entity recognition), the chance of exploiting the long tail of less common artifacts remains long underrepresented. On the other hand, an exposition of the trivial artifacts, e.g., a resolution of the correct pronoun, when the referenced subject is already referenced in the unmasked segment, occurs commonly.

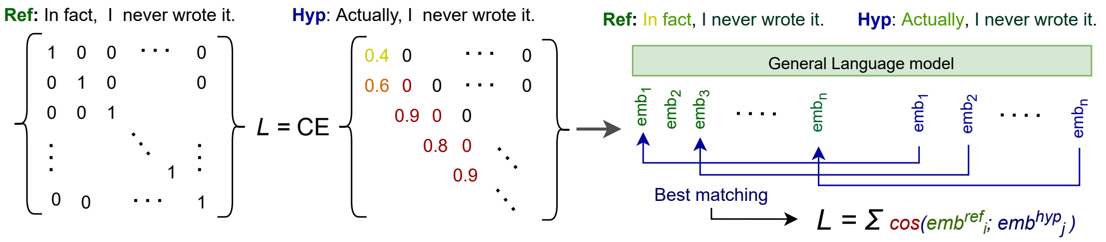
Ref: In fact, I never wrote it.
Hyp: Actually, I never wrote it.
L = CE
0.4
0.6
0.9
0.8
0.9
Ref: In fact, I never wrote it.
Hyp: Actually, I never wrote it.
General Language model
emb1
emb2
emb3
embn
emb1
emb2
embn
Best matching
L = Σ cos(emb^ref_i; emb^hyp_j)

Fig. 2: Instead of using the cross-entropy (CE) exact-matching objectives, we propose to elaborate into using "soft" objectives, able to distinguish between the different levels of *inexact* matching. As an example, we propose to compute a loss of sequence-to-sequence training objective *irrespective* of the relative ordering of the tokens in the reference and hypothesis. Similar to the evaluation of Zhang [46], the objective would first find the best-possible matching between two sets of tokens based on the token embeddings, and only then computes the value of the loss as a sum of minimal possible distances of every token in the hypothesis. Note that such objective is still differentiable on a sequence level.

> *We should ask whether the commonly-used objectives expose the full variety of the learned task in an* efficient *manner, as the efficiency will always be a qualitative bottleneck for many low-resource or domain-specific applications.*

The inefficiency, as well as the potential of objectives improvement, is exploited by the approach of ELECTRA model [7]. ELECTRA uses a simpler language model to exchange some words in the pre-training corpora. The language model is trained to distinguish the synthetically-exchanged tokens in the token classification objective instead of using the classic MLM objective. Using this approach, authors report 30x speedup of convergence while reaching very similar performance on a set of GLUE [41] tasks.

Another significant work in this direction is the one of Szegedy et al. [36], which introduces commonly-used Label smoothing nowadays. In this training strategy, the "true" distribution of labels to which the model's loss is computed is not discrete, i.e., in the form of a one-hot vector of a size of several classes $|C|$. Instead, it has a form of a vector with the values of $\frac{\epsilon}{|C|}$ on the positions of non-expected category, and a value of $1 - \epsilon$ on the true-category position, where ϵ remains a free parameter, usually set in $\langle 0.05; 2 \rangle$. Such smoothing of the objective is shown to minimize in-domain test error [36] and can improve model generalization ability [6].

These results motivate us to revisit the commonly-used objectives, where a speed of convergence and generalization can be defining factors of model's end quality, for instance, in a neural machine translation of under-resourced languages or non-canonical domains [32].

We follow with a brief motivational introduction to the problem and a proposition of one specific machine translation objective following the call for softer objectives. The approach is also summarised in Figure 2.

The standard neural machine translation objective is to minimise the cross-entropy (CE) loss between an *expected* pseudo-probabilistic distribution over model's vocabulary, for each token P_i^E given by the model, and a *true* token $y_i' \in Y^T$ given by a set of *reference translations*. P^E is *conditioned* by both the tokens of the source sequence $x_{1\dots n}$ and the *previous* tokens $y_{1\dots n-1}$. The cross-entropy token-level loss $\mathcal{L}$ is then defined as:

$$\mathcal{L}(X, y'_{1\dots n}) = \sum_{i=1}^{|n|} CE(P_i^E(X, y'_{1\dots i-1}), y_i') \,.$$

Utilising $\mathcal{L}$ in the training process, the model is trained to *predict* all P_i^E any unknown X, but compared to the training, on inference, P_i^E is conditioned by the *previously-predicted* tokens $y_{1\dots i-1}$ instead of the tokens of the reference $y'_{1\dots i-1}$.

Among other aspects, $\mathcal{L}$ implies that if the model generates one extra token or omits one token at the beginning of generation, all the subsequently-generated tokens will be sanctioned the same as if the model generated the remaining output randomly. A similar penalization is backpropagated if the model fully *paraphrases* the reference. Such a loss origin might arguably cause the model to *overfit* the syntax of the training domain, or might be the reason why the other objectives, such as Denoising [21] significantly enhance a *fluency* of output, as compared to the described Causal language modeling, as in GPT [31].

One of the simple approaches to eliminate this problem is to start with picking a reference token y_j' which is *best-matching* to the evaluated x_i. A separate, discriminative language model can provide the representations of the matched tokens, similarly to [7]. The pairwise distance of the tokens can be estimated using the max-product approach as proposed in BERTScore [47], using the many-to-many matching utilizing Wasserstein distance [20], or using any other differentiable token-level distance measure.

4.3 Objectives Utilizing Generalization Measures

> "What we demand is vigilance and attention on the part of the master and the pupils." J. A. Comenius [8]

A relatively specific direction towards higher robustness of language models is outlined by the works utilizing the approximations of measures that correlate well with empirical out-of-distribution performance. These works overviews Section 3.2. Even though some of the incorporated measures do not consistently correlate to out-of-distribution performance, from a limited number of the referenced applications, it seems that the model is always able to utilize the adjacent information efficiently.

Task-specific training objectives can be extended with an additive component, in a form outlined in Equation (1).

$$\mathcal{L}(M) = (1-\alpha)\mathcal{L}_{Obj}(M) + \alpha\mathcal{L}_{Meas}(M) \qquad (1)$$

> *To enhance model's distributional robustness, a task-specific training objective $\mathcal{L}_{Obj}$ can be additively complemented with a differentiable instance of the generalization measure $\mathcal{L}_{Meas}$.*

The measures that highly correlate with out-of-distribution accuracy of the model can be utilized to effectively regularize the final objective $\mathcal{L}$ favouring the property associated with distributional robustness. We overview some of such generalization measures in Section 3.1.

We identify two challenges in training objective design. The first one is in designing a differentiable and computationally-feasible approximation of the generalization measure. Foret [12] demonstrates that the valuation of sharpness of the parametrized space requires a valuation for all the inputs of the parametrized, application-dependent distance. It is not clear if a similar representative valuation would be feasible in the NLP domain.

The second challenge lies in designing the evaluation measures well-correlated with out-of-distribution performance and their representative evaluation. For example, Dziugaite [11] shows that the measures that correlate highly in one context might correlate poorly under different shifts. A representative evaluation of the generalization ability of the measure requires identification of all valid biases, which is not feasible, implying that the evaluation of generalization measures will remain merely the point estimates of unknown shift.

We can still escape this uncertainty in designing the generalization measures reflecting the features of the problem, which we intuitively consider to be invariant to the data domain or problem on hand. Such features could, for instance, reflect the shared linguistic properties of the natural language.

5 Conclusion

This work outlines the three directions of addressing the unwanted data biases of language models, which is an extensively reported problem inherently raised from the expressivity of the deep models.

We aim to motivate the research in these three directions, providing a shared framework and referencing the current work showing initial, promising results.

We acknowledge that there might be multiple unforeseen obstacles in any proposed directions that will only identify in practice. We argue that any contribution towards more robust language models has immediate implications for most of the applications in the NLP field. Many of the commonly-used solutions already rely on transformers and can even be seen to expose unknown, notorious biases, as shown, e.g., in [35]. At the same time, a limited extrapolation ability of the models remains a blocker for applying modern NLP in more niche domains, where little annotated data is available due to the size or audience background.

References

1. Bartlett, P.L., Foster, D.J., Telgarsky, M.: Spectrally-Normalized Margin Bounds for Neural Networks. In: Proc. of the 31st International Conference on Neural Infor-

mation Processing Systems. pp. 6241–6250. NIPS '17, Curran Associates Inc., USA (2017). https://doi.org/10.5555/3295222.3295372

2. Belinkov, Y., Bisk, Y.: Synthetic and Natural Noise Both Break Neural Machine Translation. CoRR **abs/1711.02173v2** (2018), https://arxiv.org/abs/1711.02173v2
3. Bengio, Y., Deleu, T., Rahaman, N., Ke, N.R., Lachapelle, S., Bilaniuk, O., Goyal, A., Pal, C.: A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. In: Proc. of International Conference on Learning Representations (2020), https://openreview.net/forum?id=ryxWIgBFPS
4. Berard, A., Calapodescu, I., Dymetman, M., Roux, C., Meunier, J.L., Nikoulina, V.: Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In: Proc. of the 3rd Workshop on Neural Generation and Translation. pp. 168–176. ACL, Hong Kong (Nov 2019). https://doi.org/10.18653/v1/D19-5617
5. Chadha, A., Andreopoulos, Y.: Improving Adversarial Discriminative Domain Adaptation. CoRR **abs/1809.03625v3** (2018), https://arxiv.org/abs/1809.03625v3
6. Chen, B., Ziyin, L., Wang, Z., Liang, P.P.: An Investigation of how Label Smoothing Affects Generalization. CoRR **abs/2010.12648** (2020), https://arxiv.org/abs/2010.12648v1
7. Clark, K., Luong, M.T., Le, Q.V., Manning, C.D.: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. CoRR **abs/2003.10555v1** (2020), https://arxiv.org/abs/2003.10555v1
8. Comenius, J.A.: Didáctica magna. Amsterdam (1649), https://webspace.ship.edu/cgboer/comenius.html, The Great Didactic, translated by M. W. Keatinge 1896
9. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale. CoRR **abs/1911.02116v2** (2020), https://arxiv.org/abs/1911.02116v2
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805v2** (2018), https://arXiv.org/abs/1810.04805v2
11. Dziugaite, G.K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I., Roy, D.M.: In Search of Robust Measures of Generalization. CoRR **abs/2010.11924v2** (2021), https://arxiv.org/abs/2010.11924v2
12. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-Aware Minimization for Efficiently Improving Generalization. CoRR **abs/2010.01412v1** (2021), https://arxiv.org/abs/2010.01412v1
13. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Proc. of the 58th Annual Meeting of the ACL. pp. 8342–8360. ACL (Jul 2020), https://aclanthology.org/2020.acl-main.740.pdf
14. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation Artifacts in Natural Language Inference Data. In: Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). pp. 107–112. ACL, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-2017
15. Hinton, G.E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation **14**(8), 1771–1800 (Aug 2002). https://doi.org/10.1162/089976602760128018
16. Iyer, S., Dandekar, N., Csernai, K.: First Quora Dataset Release: Question Pairs (2017), https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

17. Jiang, J., Zhai, C.: Instance Weighting for Domain Adaptation in NLP. In: Proc. of the 45th Annual Meeting of the ACL. pp. 264–271. ACL, Prague, Czech Republic (Jun 2007), `https://aclanthology.org/P07-1034`
18. Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., Bengio, S.: Fantastic Generalization Measures and Where to Find Them. CoRR **abs/1912.02178v1** (2020), `https://arxiv.org/abs/1912.02178v1`
19. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. CoRR **abs/1609.04836v1** (2017), `https://arxiv.org/abs/1609.04836v1`
20. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From Word Embeddings To Document Distances. In: Bach, F., Blei, D. (eds.) Proc. of International Conference on Machine Learning. vol. 37, pp. 957–966. PMLR, Lille, France (Jul 2015), `http://proceedings.mlr.press/v37/kusnerb15.html`
21. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proc. of the 58th Annual Meeting of the ACL. pp. 7871–7880 (2020), `https://aclanthology.org/2020.acl-main.703.pdf`
22. Lin, S., Tsai, C., Chien, L., Chen, K., Lee, L.: Chinese language model adaptation based on document classification and multiple domain-specific language models. In: Kokkinakis, G., Fakotakis, N., Dermatas, E. (eds.) Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997. ISCA, Rhodes, Greece (Sep 1997), `http://www.isca-speech.org/archive/eurospeech_1997/e97_1463.html`
23. Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., Bachem, O.: Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. CoRR **1811.12359v4** (2019), `https://arXiv.org/abs/1811.12359v4`
24. McCoy, T., Pavlick, E., Linzen, T.: Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In: Proc. of the 57th Annual Meeting of the ACL. pp. 3428–3448. ACL, Florence, Italy (Jul 2019). `https://doi.org/10.18653/v1/P19-1334`
25. Moore, R.C., Lewis, W.: Intelligent Selection of Language Model Training Data. In: Proc. of the ACL Conference. pp. 220–224. ACL, Uppsala, Sweden (Jul 2010), `https://aclanthology.org/P10-2041`
26. Nie, Y., Wang, Y., Bansal, M.: Analyzing Compositionality-Sensitivity of NLI Models. CoRR **abs/1811.07033v1** (2019), `https://arxiv.org/abs/1811.07033v1`
27. Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., Kiela, D.: Adversarial NLI: A new benchmark for natural language understanding. In: Proc. of the 58th Annual Meeting of the ACL. pp. 4885–4901. ACL (Jul 2020). `https://doi.org/10.18653/v1/2020.acl-main.441`
28. Pitas, K., Davies, M.E., Vandergheynst, P.: PAC-Bayesian Margin Bounds for Convolutional Neural Networks. CoRR **abs/1801.00171v2** (2018), `https://arxiv.org/abs/1801.00171v2`
29. Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., Van Durme, B.: Hypothesis Only Baselines in Natural Language Inference. In: Proc. of the Seventh Joint Conference on Lexical and Computational Semantics. pp. 180–191. ACL, New Orleans, USA (Jun 2018). `https://doi.org/10.18653/v1/S18-2023`, `https://aclanthology.org/S18-2023`
30. Popel, M., Tomková, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., Žabokrtský, Z.: Transforming machine translation: a deep learning system reaches news transla-

tion quality comparable to human professionals. Nature Communications **11**(4381) (2020). https://doi.org/10.1038/s41467-020-18073-9

31. Radford, A., Narasimhan, K.: Improving Language Understanding by Generative Pre-Training (2018), https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
32. Ramponi, A., Plank, B.: Neural unsupervised domain adaptation in NLP—A survey. In: Proc. of the 28th International Conference on Computational Linguistics. pp. 6838–6855. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.603
33. Sanh, V., Wolf, T., Belinkov, Y., Rush, A.M.: Learning from others' mistakes: Avoiding dataset biases without modeling them. CoRR **abs/2012.01300v1** (2021), https://arxiv.org/abs/2012.01300v1
34. Shah, D.J., Lei, T., Moschitti, A., Romeo, S., Nakov, P.: Adversarial Domain Adaptation for Duplicate Question Detection. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proc. of the 2018 Conference EMNLP. pp. 1056–1063. ACL (2018). https://doi.org/10.18653/v1/d18-1131
35. Štefánik, M., Novotný, V., Sojka, P.: RegEMT: Regressive Ensemble for Machine Translation Quality Evaluation. In: Proc. of the Sixth Conference on Machine Translation (WMT). pp. 1046–1053. ACL (Nov 2021), https://www.statmt.org/wmt21/pdf/2021.wmt-1.112.pdf, poster also available: https://mir.fi.muni.cz/posters/emnlp-2021-regemt.pdf
36. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. In: IEEE Conf. CVPR. pp. 2818–2826. IEEE, Los Alamitos, USA (Jun 2016). https://doi.org/10.1109/CVPR.2016.308
37. Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: Proc. of the 57th Annual Meeting of the ACL. pp. 4593–4601. ACL, Florence, Italy (Jul 2019). https://aclanthology.org/P19-1452
38. Tu, L., Lalwani, G., Gella, S., He, H.: An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. Transactions of the ACL **8**, 621–633 (Oct 2020). https://doi.org/10.1162/tacl_a_00335
39. Utama, P.A., Moosavi, N.S., Gurevych, I.: Towards Debiasing NLU Models from Unknown Biases. In: Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7597–7610. ACL, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.613
40. Valiant, L.G.: A Theory of the Learnable. In: Proc. of the Sixteenth Annual ACM Symposium on Theory of Computing. pp. 436–445. STOC '84, ACM, New York, USA (1984), https://doi.org/10.1145/800057.808710
41. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In: Proc. of the 2018 EMNLP Workshop BlackboxNLP. pp. 353–355. ACL, Brussels, Belgium (Nov 2018), https://aclanthology.org/W18-5446
42. Williams, A., Nangia, N., Bowman, S.: A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In: Proc. of the 2018 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long Papers). pp. 1112–1122. ACL (2018), https://aclweb.org/anthology/N18-1101
43. Wu, M., Moosavi, N., Rücklé, A., Gurevych, I.: Improving QA Generalization by Concurrent Modeling of Multiple Biases. CoRR (2020), https://arxiv.org/abs/2010.03338v1
44. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised Data Augmentation. CoRR **abs/1904.12848v1** (2019), https://arXiv.org/abs/1904.12848v1

45. Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In: 33rd Annual Meeting of the ACL. pp. 189–196. ACL, Cambridge, Massachusetts, USA (Jun 1995). `https://aclanthology.org/P95-1026`
46. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with BERT. CoRR **abs/1904.09675v3** (2019), `https://arxiv.org/abs/1904.09675v3`
47. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. In: Proc. of International Conference on Learning Representations (2020), `https://openreview.net/forum?id=SkeHuCVFDr`
48. Zhang, Y., Baldridge, J., He, L.: PAWS: Paraphrase Adversaries from Word Scrambling. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proc. of the 2019 Conf. NAACL-HLT. pp. 1298–1308. ACL, Minneapolis, USA (Jun 2019). `https://doi.org/10.18653/v1/n19-1131`
49. Zhou, Z.H., Li, M.: Tri-training: exploiting unlabeled data using three classifiers. IEEE Transactions on Knowledge and Data Engineering **17**(11), 1529–1541 (2005). `https://doi.org/10.1109/TKDE.2005.186`