ADBCMM : Acronym Disambiguation by Building Counterfactuals and Multilingual Mixing

Yixuan Weng[1], Fei Xia[1,2], Bin Li[3], Xiusheng Huang[1,2] and Shizhu He[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation,Chinese Academy Sciences, Beijing, 100190, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100190, China
[2] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy Sciences, Beijing, 100190, China
[4]College of Electrical and Information Engineering, Hunan Universit

Abstract

Scientific documents often contain a large number of acronyms. Disambiguation of these acronyms will help researchers better understand the meaning of vocabulary in the documents. In the past, thanks to large amounts of data from English literature, acronym task was mainly applied in English literature. However, for other low-resource languages, it's training data is scarce, so the generalization performance of the model is poor. To address the above issue, this paper proposes a new method for acronym disambiguation, named as ADBCMM, which can significantly improve the performance of low-resource languages by building counterfactuals and multilingual mixing. Specifically, by balancing data bias in low-resource langauge, ADBCMM will able to improve the test performance outside the data set. In SDU@AAAI-22 - Shared Task 2: Acronym Disambiguation, the proposed method won first place in French and Spanish. You can repeat our results here https://github.com/WENGSYX/ADBCMM.

Keywords

Acronym Disambiguation, Counterfactuals, Low resource, Minority Language

1. Introduction

The exchanges between countries become closer with the progress of globalization. As countries began to communicate more politically, economically and academically, language understanding became a new challenge. Acronyms often appear in the scientific documents of different countries. Compared to English, acronyms are more challenging to understand in other languages. Acronyms will become a barrier for researchers to read scientific literature and affect exchanges and cooperation between countries.

Acronym disambiguation refers to the problem of solving the large number of acronym differences in the text, which need to find the correct interpretation. For these acronyms, we need to find the correct one in the current context from the dictionary. For example, in "The traditional Chinese sentences are transferred into SC", "SC" means "simplified Chinese" rather than "System Combination". It is difficult for some people who are not familiar with a language to understand related acronyms. So we need to distinguish abbreviations, which is a challenging task.

Figure 1: Differences and challenges between English and other (such as French) phrases in the acronym disambiguation. Red means wrong, green means right. Acronyms in English are often first letter acronyms, but not in other languages.

In the datasets, 30,237 data in the four fields of English (science), English (legal), French and Spanish were given. Each datapoint contains a sentence and there will appear a word that is the first letter abbreviated. The task hopes to find the most suitable form of an extension for the first letter abbreviation.

In the past, researchers have tried to solve AD problems by means of character extraction [1], word embedding [2], and deep learning [3]. Over the last few years, the BERT [4] model has emerged, which adopts a method of pre-training in a large language library. Many studies have shown that these pre-training models (PTMs) have gained a wealth of generic characteristics. Recently, They

SDU@AAAI-22: Workshop on Scientific Document Understanding, co-located with AAAI 2022. 2022 Vancouver, Canada.
wengsyx@gmail.com (Y. Weng); xiafei2020@ia.ac.cn (F. Xia); Mlibincn@hnu.edu.cn (B. Li); huangxiusheng2020@ia.ac.cn (X. Huang); shizhu.he@nlpr.ia.ac.cn (S. He)
https://wengsyx.github.io/ (Y. Weng)

[5, 6] have achieved remarkable effects using the BERT model in AD tasks.

However, these methods do not work well in other languages. So we used the following methods to further enhance the model's out-of-data test performance to help better researchers understand and communicate multilingual multi-domain scientific documents.

- A simple ADBCMM approach was proposed to use other language data as counterfacts datasets in AD tasks, solving the model bias.
- We tried to use the Multiple-Choice Model framework to make the model more focused on word-to-word comparisons to help the model better understand the first letter abbreviation.
- Our results achieved SOTA effects in both the French and Spanish of the AD dataset, showing outstanding performance, surpassing all other baselines methods.

2. Related Work

In this section, we will introduce the Acronym Disambiguation datasets[7] and how to solve the Acronym Disambiguation tasks [8] in English scenarios in the past, while introducing the difficulties of the Acronym Disambiguation tasks in other languages.

2.1. AD dataset

Table 1
Specific number of the Acronym Disambiguation datasets. Including the Acronym Disambiguation tasks for 4 different fields. The total number of data sets is not more than 10,000.

Data	En(Lagel)	En(Sci)	French	Spanish
Train	2949	7532	7851	6267
Dev	385	894	909	818
Test	383	574	813	862
Total	3717	9000	9573	7947

In this AD task, the abbreviation appears in scientific documents in English and other languages. AD datasets provide datasets in French and Spanish in addition to English. Each data gives a dictionary, and each language split has its test set with acronyms not appearing in their training set.

2.2. Previous work

In the AD of SDU@AAAI-21, the teams presented their methodologies and submitted a total of 10 papers. Those papers included some excellent projects.

Pan [5] trained a Binary Classification Model incorporating BERT and several training strategies. His program includes dynamic adverse sample selection, task adaptive pretraining, adversarial training [9] and pseudo labelling in his paper. This model achieved its first achievement.

Zhong [6] believes that different pre-training models store knowledge in different fields, and better results can be achieved through model integration. He proposed a Hierarchical Dual-path BERT method to capture general and professional field language, while using RoBERTa and SciBERT to perceive and predict text. He eventually reached a 93.73% F1 value in the SciAD datasets.

2.3. Difficulty with multilingual data

In the AD of SDU@AAAI-22, the organizers released AD datasets covering French and Spanish, which have the following difficulties compared to the English environment:

- In Figure 1, we can find that the extension of other languages does not necessarily contain an acronym of the first letter, and it isn't easy to match directly through the rules.
- Other languages lack PLMs trained in scientific language.
- In Table 1, the number of datasets in French and Spanish is small. Training models are prone to bias and over-adaptation.

3. Methods

In this section, we will describe the framework for the overall model, as well as a range of methods for AD datasets for other languages, including ADBCMM, In-Trust-loss [10], Child-Tuning [11] and R-Drop[12].

3.1. The model framework

We use the Multiple-Choice model framework, which is different from the Binary Classification Model used by Pan [5].

The Multiple-Choice model [13] refers to adding a classifier to the end output of the BERT model. Each sentence has only a single output value to represent the probability of this option.

In Figure 2, when we use the Multiple-Choice model, each batch will enter all the possible options in the same set during the training. If the word in the dictionary is insufficient, we use "Padding" for filling, eventually at the output end for softmax classification and calculation of losses.

Thus, we can more accurately derive the probability of each option by comparing methods. Compared with Binary Classification Model, Multiple-Choice model capturing more semantic characteristics and make the model

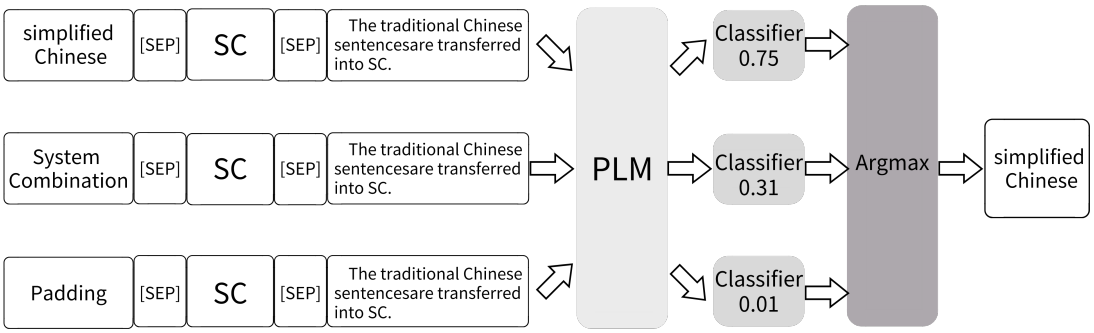

Figure 2: Multiple-Choice Model

more comprehensively trained and predicted on differences, rather than the error interference model caused by the dynamic construction of negative samples.

3.2. ADBCMM

PLM has achieved excellent results in many NLP tasks, but the potential bias in training data can harm out-of-data testing performance. Counterfactually augmented datasets is a recent solution [14]. But if it takes a lot of human resources and money to build counterfactual samples by man, this approach is not realistic.

We found many homonyms samples by analyzing erroneous samples on dev datasets. We think these samples errors are mainly due to model bias, over-training leads to over-adaptation seriously, and data set performance is poor. That's why we used different language markup information to use other language samples as new counterfactual samples after being modified.

In Figure 3, the training process is like a pyramid. We first train using data in multiple languages, and then we do secondary training in a single language based on pre-training.

Why continue training with single-language materials after multilingual mixed training instead of testing directly after multilingual Counterfacts datasets training? Because in our experiment, with the addition of more language samples, the models may become overwhelming. Even though French, English and Spanish belong to the Indo-European language family, they all have unique language properties, syntax and vocabulary. This would be a noise interference for different languages. Models may ignore semantic characteristics that are unique to a particular language and prefer to learn more common ones.

Our ADBCMM approach can also be further extended to translation, Ner, conversation generation and other tasks. The ADBCMM approach helps address biases caused by insufficient data in small-language environments.

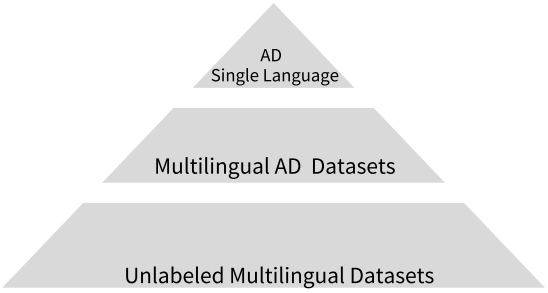

Figure 3: Training Process

3.3. Child-Tuning

Because AD data sets are smaller and can easily be learned, resulting in the model's poor centralized generalization capacity during testing. We used the Child-Tuning method proposed to address this discrepancy. The Child-Tuning [11] strategy only updates the corresponding Child Network when the parameters are updated backwards, without adjusting all the parameters. At the end of the first epoch, we compare the model's parameters with the original parameters to find out the greatest weight of the change, and in the subsequent we only update the parameters of this section. This approach like the reverse Dropout [15], it can bring performance improvements to our models.

Table 2
Experimental results in French and Spanish AD datasets. BETO is a Spanish pre-training model, tested only on Spanish data in AD; Flaubert-base-cased is a French pre-training model, tested only on French data in AD; mDeberta is a multi-language pre-training model, we test in both French and Spanish. Additionally, methods including "ADBCMM", "Child-Tuning", "R-Drop" and "Alls" are fine-tuned on mDeberta models, "Alls" refers to using all of the above methods. In addition to "Finally in Test", we test the results of the Dev series. "Finally in Test" also uses model fusion to improve our performance.

Language	**French**			**Spanish**		
Model/Method	Precision	Recall	Macro F1	Precision	Recall	Macro F1
BETO	N/A	N/A	N/A	0.8063	0.7510	0.7777
Flaubert-base-cased	0.7796	0.6786	0.7256	N/A	N/A	N/A
mDeberta-v3-base	0.7244	0.6001	0.6564	0.7176	0.6491	0.6816
+ ADBCMM	0.8087	0.7213	0.7625	0.8558	0.8236	0.8394
+ Child-Tuning	0.7438	0.6232	0.6782	0.7512	0.6834	0.7157
+ R-Drop	0.7467	0.6337	0.6856	0.7492	0.7019	0.7248
ALLs	**0.8423**	**0.7712**	**0.8052**	**0.8859**	**0.8352**	**0.8598**
Finally in Test	**0.8942**	**0.7934**	**0.8408**	**0.9107**	**0.8514**	**0.8801**

3.4. R-Drop

In the R-Drop work, the authors used the model to open Dropout during the training and then made two inputs, so the results of the two inputs would not be the same because the model opened Dropout. In addition to calculating the loss of label information, the Kullback-Leibler divergence was also calculated between the same two inputs but different outputs. This R-Drop method can play the role of normalizing and increasing robustness. In our experiment, R-Drop improved greater performance.

4. Experimental Setting

This section will subsequently present our Baseline, experimental models, experimental settings, control of variables experiment.

4.1. Baseline

For both French and Spanish languages, we used Flaubert-base-cased [16] models and BETO [17] cased models respectively. These models are Bidirectional Encoder Representations from Transformers [4], and the size is both bases. These models have a lot of Masked Language Model (MLM) [18] training in the related large single-language repository and have state-of-the-art(SOTA) results in the related languages. These pre-trained models can better capture the semantic information of words.

But there is no additional training, so the two models still need to fine-tune AD data centralization to solve the Acronym Disambiguation tasks. We will add a classification layer behind these models, and then the models become Multiple-Choice Models. We trained the models in a single language. Their results will be used as our Baseline, and the results of other models will be compared with them.

4.2. Model

To better adapt to the ADBCMM method, we used the DeBERTa model [19, 20] for pre-training in the multilingual repository CC100 [21] . The authors of DeBERTa replaced the MLM objective with the RTD (Replaced Token Detection) intent introduced by ELECTRA[22] for pre-training.

Specifically, we used the mdeberta-v3-base[1] model in the experiment, with a total of 280M and containing 250,000 tokens. MDeberta supports 100 languages in 100 countries, including English, French and Spanish.

Of course, to ensure that the ADBCMM method rather than the mDeberta model brought us practical performance enhancements, we also used mDeberta only in French or Spanish as a contrast experiment.

4.3. Parameters Setup

We used three pre-training models, including Flaubert, BETO and mDeberta, for a total of 15 training sessions. We use argmax to choose the maximum of all values as the final result for the word to be selected.

In all the experiments, we set 16 epochs and decided to use the 1e-5 learning rate (we used warmup simultaneously) with Pytorch[23]. We put gradient decrease 1e-5 and batch size 1 (each batch contains 14 different options). we employ the AdamW optimizer [24] and use the hugging-face[2] [13] framework. We only use the first 300 tokens for each sample. On a Intel 10900K server with 128G memory, we used a 24G NVIDIA 3090 GPU to train our model.

[1]You can go to https://huggingface.co/microsoft/mdeberta-v3-base download model

[2]https://github.com/huggingface/transformers

Table 3
SDU@AAAI ranks of the Acronym Disambiguation tasks in French and Spanish

	French			Spanish		
Ranked	Precision	Recall	Macro F1	Precision	Recall	Macro F1
Rank1(Ours)	**0.89**	**0.79**	**0.84**	**0.91**	**0.85**	**0.88**
Rank2	0.85	0.73	0.78	0.88	0.79	0.83
Rank3	0.81	0.72	0.76	0.86	0.80	0.83
Rank4	0.76	0.70	0.73	0.83	0.80	0.81
Rank5	0.73	0.64	0.68	0.86	0.77	0.81

4.4. Assessment of indicators

In AD tasks, Macro F1 was used as an assessment indicator by calculating the accuracy and recall rate of the final result.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2PrecisionRecall}{Precision + Recall}$$

$$MacroF1 = \frac{\sum_{i=1}^{n} F1_i}{n}$$

n means that the higher the total number of categories, accuracy, recall rate, and MacroF1. The higher the F1 method, the better the performance.[3]

5. Results

In Table 2, we can find that under the same conditions, mDeberta performs less well in French than in Flaubert-base-cased, and less well in Spanish than in BETO. We speculate that because mDeberta uses a large number of data in different languages during the pre-training phase. Still, after spinning into other languages, due to the further side focus, it may not necessarily accurately record the semantic characteristics of a single language so that the actual performance will be slightly worse compared to BETO and Flaubert. They have been pre-trained only in a single language.

Both Child-Tuning and R-Drop showed excellent performance in English and Spanish, bringing a 3-5% F1 boost to our model. But compared to the ADBCMM method, they were still slightly underperforming. Our ADBCMM method brought more than 10% performance boost directly to our mDeberta model. This is indeed incredible. To ensure the repetitiveness of the experiment, we repeated three experiments. The mDeberta models using the ADBCMM method were compared to their mDeberta model F1 performance over 10% in these three experiments.

We think that ADBCMM can significantly boost our models because of the reliable Counterfacts datasets. First, they can match upstream and downstream training data; second, counterfacts datasets can reduce the model's bias, learning from more text data to more relevant information with Acronym Disambiguation tasks; third, even if the datasets are collected from different languages or fields, but they are scientific documents, so the general language training mDeberta model can learn the syntax characteristics of scientific documents in more scientific documents and further improve performance.

Finally, we followed ADBCMM-based methods and achieved SOTA scores in both SDU@AAAI's French and Spanish. In Acronym Disambiguation tasks [8], our methods of Precision, Recall and Macro F1 are SOTA. Remarkably, our approach leads us to the second F1 score of 5% - 6%.

6. Conclusion

In this article, we mainly talk about how to use ADBCMM in the Acronym Disambiguation tasks at SDU@AAAI-22 and compare it with other Models or Methods to yield SOTA. We used a straightforward method to build counterfacts datasets in ADBCMM. We directly use other language datasets for training and secondary Fine-Tune in their language, which gives our models a remarkable effect. After combining the Multiple-Choice Model, Child-Tuning, R-Drop and other methods, our approach leads ahead of all different systems. Apparently, in multilingual data aggregation, simply using other languages as counterfacts datasets can improve performance. At the same time, our work provides practical help for researchers to understand scientific documentation better.

[3]Below is the specific meaning of the formula. TP: The prediction is correct and the sample is correct. FP: The prediction is wrong and the sample is correct. FN: The predicting is correct and the sample is wrong.

7. Acknowledgement

The work is supported by the National Key Research and Development Program of China (2020AAA0106400) and the National Natural Science Foundation of China (61922085, 61976211). The work is also supported by the Beijing Academy of Artificial Intelligence (BAAI2019QN0301), the Key Research Program of the Chinese Academy of Sciences under Grant (ZDBS-SSW-JSC006), the independent research project of the National Laboratory of Pattern Recognition, China and the Youth Innovation Promotion Association CAS, China.

8. Online Resources

- GitHub,https://github.com/WENGSYX/ADBCMM

References

[1] Y. Li, B. Zhao, A. Fuxman, F. Tao, Guess me if you can: Acronym disambiguation for enterprises, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 1308–1317. URL: https://aclanthology.org/P18-1121. doi:10.18653/v1/P18-1121.

[2] J. Charbonnier, C. Wartena, Using word embeddings for unsupervised acronym disambiguation, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 2610–2619. URL: https://aclanthology.org/C18-1221.

[3] Q. Jin, J. Liu, X. Lu, Deep contextualized biomedical abbreviation expansion, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 88–96. URL: https://aclanthology.org/W19-5010. doi:10.18653/v1/W19-5010.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[5] C. Pan, B. Song, S. Wang, Z. Luo, Bert-based acronym disambiguation with multiple training strategies, ArXiv abs/2103.00488 (2021).

[6] Q. Zhong, G. Zeng, D. Zhu, Y. Zhang, W. Lin, B. Chen, J. Tang, Leveraging domain agnostic and specific knowledge for acronym disambiguation, ArXiv abs/2107.00316 (2021).

[7] S. Y. R. J. F. D. T. H. N. Amir Pouran Ben Veyseh, Nicole Meister, MACRONYM: A Large-Scale Dataset for Multilingual and Multi-Domain Acronym Extraction, in: arXiv, 2022.

[8] S. Y. R. J. F. D. T. H. N. Amir Pouran Ben Veyseh, Nicole Meister, Multilingual Acronym Extraction and Disambiguation Shared Tasks at SDU 2022, in: Proceedings of SDU@AAAI-22, 2022.

[9] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, CoRR abs/1412.6572 (2015).

[10] X. Huang, Y. Chen, S. Wu, J. Zhao, Y. Xie, W. Sun, Named entity recognition via noise aware training mechanism with data filter, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 4791–4803. URL: https://aclanthology.org/2021.findings-acl.423. doi:10.18653/v1/2021.findings-acl.423.

[11] R. Xu, F. Luo, Z. Zhang, C. Tan, B. Chang, S. Huang, F. Huang, Raise a child in large language model: Towards effective and generalizable fine-tuning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 9514–9528. URL: https://aclanthology.org/2021.emnlp-main.749.

[12] X. Liang, L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, T.-Y. Liu, R-drop: Regularized dropout for neural networks, in: NeurIPS, 2021.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[14] D. Kaushik, A. Setlur, E. H. Hovy, Z. C. Lipton, Explaining the efficacy of counterfactually augmented data, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=HHiiQKWsOcV.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learn-

ing Research 15 (2014) 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[16] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, D. Schwab, Flaubert: Unsupervised language model pre-training for french, in: Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2479–2490. URL: https://www.aclweb.org/anthology/2020.lrec-1.302.

[17] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[18] W. L. Taylor, "cloze procedure": A new tool for measuring readability, Journalism Quarterly 30 (1953) 415–433. URL: https://doi.org/10.1177/107769905303000401. doi:10.1177/107769905303000401. arXiv:https://doi.org/10.1177/107769905303000401.

[19] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[20] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[21] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://www.aclweb.org/anthology/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[22] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, in: ICLR, 2020. URL: https://openreview.net/pdf?id=r1xMH1BtvB.

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[24] I. Loshchilov, F. Hutter, Fixing weight decay regularization in adam, ArXiv abs/1711.05101 (2017).