Multilingual Acronym Disambiguation with Multi-choice Classification

Xinyu Zhu[1], Chengze Yu[1], Siheng Li[1], Tian Liang[1], Cheng Yang[1] and Yujiu Yang[1]

[1]Tsinghua Shenzhen International Graduate School,Tsinghua University, Shenzhen, P. R. China

Abstract

Acronym disambiguation (AD) is a practical task that aims at matching acronyms with their expansion alternatives. The significance of this task is to alleviate the problem of acronym abuse, which is a common phenomenon in the scientific domain. Although it simplifies the expression, great obstacles are brought to readers outside the field. In this paper, we introduce a new method for acronym disambiguation, which is different from the previous binary classification model. We argue that treating the task as a multi-choice problem and incorporating negative expansion sampling can more effectively capture the relation between the acronym and its corresponding expansion. Experiments show our model obtains significant improvement over the baseline and other methods.

1. Introduction

In recent years, with the continuous progress in the scientific domain, the amount of technical vocabulary is growing at an incredible rate. In order to simplify the expression of technical contexts, acronyms are frequently used to avoid repeating the complex long-form of scientific phrases.

An acronym, a short form of a phrase, is generally made up of the initial character of its corresponding expansion. For instance, in sentence "*All systems use their* **IP** *address to introduce themselves to the network.*", **IP** here is the abbreviation of "*Internet Protocol*".

However, an acronym could be related to multiple expansions, so that the abuse of acronyms may lead to difficulties in semantic comprehension. For example, **IP** can be an acronym for "*Internet Protocol*" or "*Intellectual Property*" in different articles. The actual meaning of it depends on the context. Due to this one-to-many problem, the acronym disambiguation becomes a practical task in scientific document understanding. Properly solving this problem can benefit multiple downstream tasks such as named entity recognition [1] and relation extraction [2, 3].

Acronym disambiguation aims to find the most appropriate expansion among several candidates, and a typical sample of this task is shown in Figure 1. The input includes a sentence and an acronym dictionary, which contain the target acronym and long-form phrase alternatives, respectively. Different long forms are colored in different colors. Since the color of the actual label is blue, we also color the acronym **PC** blue to indicate the matching relationship between these two forms.

Input:

We also see that the system ability to distinguish Correct and PC answers need to be improved.

PC: -- Partially Correct -- pathway curation -- prepositional complement

Output: Partially Correct

Figure 1: An instance of acronym disambiguation

In the last year, SDU@AAAI-21 has already studied the acronym disambiguation in the scientific domain, but the language is limited to English. SDU@AAAI-22 [4] expands the corpus to English (Legal domain and Scientific domain), French, and Spanish to further explore this problem. Besides, to thoroughly test the robustness of the model, this task refers to the formulation of zero-shot tasks, which means the acronyms in the train/dev/test set are exclusive, making the task a zero-shot problem within each domain. Datasets of this task is provided by Amir Pouran Ben Veyseh [5].

Generally, the acronym disambiguation is formulated as a classification problem. In last year's competition, Pan et al. [6] trained a binary classification model. However, we find that treating the task as a multi-choice problem and incorporating negative expansion sampling can be more efficient on this year's multilingual and zero-shot corpus. Furthermore, by utilizing BERT-based pre-trained language models(PLMs), we are able to capture

zhuxy21@mails.tsinghua.edu.cn (X. Zhu);
ycz21@mails.tsinghua.edu.cn (C. Yu);
lisiheng21@mails.tsinghua.edu.cn (S. Li);
liangt21@mails.tsinghua.edu.cn (T. Liang);
yangc21@mails.tsinghua.edu.cn (C. Yang);
yang.yujiu@sz.tsinghua.edu.cn (Y. Yang)

© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

the contextual information and leverage it in the process of classification.

The main contributions of this work are as follows:

- We propose a multi-choice classification model to solve the acronym disambiguation task and demonstrate its effectiveness.
- We propose a method called Negative Expansion Sampling method to help the model distinguish the different expansion candidates more effectively.
- We utilize various BERT-based PLMs and different training strategies on the multilingual corpus and explore the optimal pre-trained model for each dataset.
- Our model achieves a competitive performance on different language tracks and finally ranks 4th in SDU@AAAI-22 competition.

2. Related Work

2.1. SDU@AAAI-21 Works

A similar competition was held last year, and participants in SDU@AAAI-21 made a significant exploration in the acronym disambiguation.

The state-of-the-artwork of this competition [6] leverages BERT [7] and RoBERTa [8] instead of training from scratch. They also utilize multiple training strategies, such as adversarial training and pseudo-labeling, to improve model performance. Second place team [9] merges domain agnostic and specific knowledge to introduce extra information for acronym disambiguation task, which achieves a comparable performance to the above one.

Besides these two teams, some other effective methods are proposed for the task. Egan and Bohannon [10] exploit distantly-supervised datasets to enhance the training data. Singh and Kumar [11] redefine the acronym disambiguation as a span prediction task. And although not the best score, Pereira et al. [12] employs SVM to obtain a highly efficient model.

2.2. Word Sense Disambiguation

Word sense disambiguation aims to identify the appropriate meaning of a word in context [13]. Compared with the acronym disambiguation, more attention is paid to this similar research field.

SensEmBERT [14] is a remarkable knowledge-based approach that enables the exploitation of knowledge in semantic representation extraction. Along with it, EWISER [15] leverages the relational information encoded in Lexical Knowledge Bases hence achieving a considerable improvement.

Confined of the limited training resources, some researchers also attempt to utilize unsupervised [16] or semi-supervised [17] approaches to alleviate this problem. This enlightening exploration obtained a satisfying result.

2.3. Pre-trained Language Models

Nowadays, it is common to utilize pre-trained language models(PLMs) as a feature extractor for the downstream NLP tasks. The prior knowledge encoded in the PLMs could significantly boost the performance and speed up the convergence of the model compared to those trained from scratch.

Typical PLMs include GPT [18] and BERT [7]. The BERT-like models have become the most popular PLMs since their emergence. Considering of the gap between different languages, we leverage RoBERTa [8], BERT-Spanish [19] and BERT-multilingual [7] on English, Spanish and French respectively. We also explore the domain-specific model on the English scientific and legal track using SciBERT [20] and LEGAL-BERT [21].

3. Proposed Method

In this section, we will formulate the acronym disambiguation task first and then give an overview of our model. Finally, we will introduce the proposed method in detail.

3.1. Task Formulation

The acronym disambiguation task is generally formulated as a classification problem. Task input includes a sentence $S = \{x_1, x_2, ..., a_p^i, ..., x_n\}$ and an expansion dictionary $D_i = \{L_1^i, L_2^i, ..., L_m^i\}$. In sentence S, token a_p^i represents the i^{th} class acronym at the p^{th}position. Dictionary D_i contains long forms L^i for the i^{th} class acronym a_p^i. The length of sentence S and dictionary D_i is $n = |S|$ and $m = |D_i|$ respectively. The model is expected to match a_i with the most appropriate long form among D_i.

3.2. Model Architecture

Figure 2 gives an overview architecture of our model. For model input, each long-form alternative is concatenated with context separated by [SEP] token. Furthermore, we add special tokens $< start >$ and $< end >$ to mark the position of the target acronym, this can help guide our model pay more attention to it.

Different from previous works [6, 9], which treat the task as a binary classification problem and every time only consider one expansion with the sentence, we prefer to take all expansions that belong to the same acronym

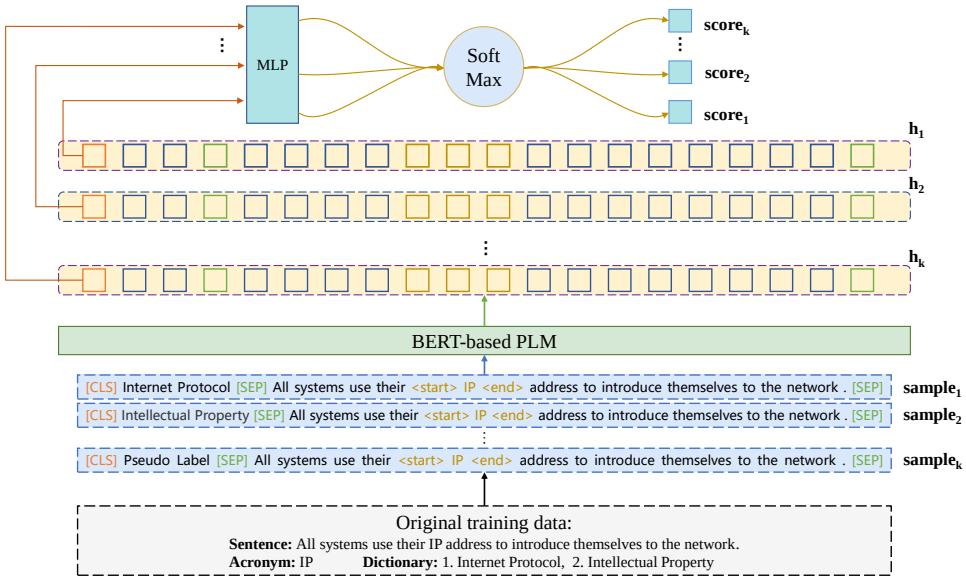

Figure 2: An overview of the proposed method. Original training data is handled to a batch of samples, and random pseudo labeled samples are used to pad the batch into constant length. With carefully selected BERT-based PLM, samples are encoded into representations, where the hidden state of [CLS] is applied to multi-choice classification.

into account at a time and model the task as a multi-choice classification problem. What's more, for each acronym a_i, we randomly sample negative expansions from other acronyms' dictionaries and add them into D_i as fake candidates to further strengthen the robustness of our model.

As a result, different from common mini-batch training that a single batch contains samples irrelevant to each other, one sample here contains different expansions from D_i concatenated with the same sentence, we randomly sample negative expansions to pad all dictionaries D_i to the same size k and we set $k = \max_i\{|D_i|\}$, so when batch size is set to bs, the true training batch size can be seen as $bs * \max_i\{|D_i|\}$. We give each long form candidate a score $f(a_i, L_k^i, s)$, where f is the function representing our model, then the probability of each long form candidate is obtained through a softmax function. We leave out the subscript index i of acronym and dictionary for simplicity.

$$P(L_k) = \frac{f(a, L_k, s)}{\sum_{n=1}^{|D|} f(a, L_n, s)} \tag{1}$$

Specifically, samples are sent to a BERT-based PLM and encoded into hidden states. Following the common practice of utilizing BERT, we leverage the representation of [CLS] token as the contextual vector, and a CLIP-Adapter [22] is used as the classifier to get each expansion's score. In addition, we incorporate two tricks, adversarial training [23] and Child-Tuning [24], to further boost model generalization. Experiments on the SDU@AAAI-22 dataset demonstrate the effectiveness of our model.

3.3. Training Strategies

3.3.1. Negative Expansion Sampling.

Due to the variable size of different acronyms' dictionaries, the way of sample construction becomes a non-trivial problem. In binary classification, each sample contains a sentence, an expansion of the acronym and a label indicating whether the expansion is the actual long form for the acronym in the sentence. For this classification, it is easy to pack several samples into a batch. However, when treating the task as multi-choice classification problem, every time we have to take all expansions of an acronym into account to get a softmax probability, but different acronyms have dictionaries with variable size, which means we have to split samples in a batch into different small groups and do softmax within the group. For example, if one batch contain 2 different sentences and the first sentence contains an acronym that has 3 possible expansions and then we concatenate each expansion with the same sentence to create 3 samples. Similarly, the second one has 5, so the actual batch size is 8, thus we have to perform softmax on the first three samples and the other five separately. This method not

Dataset	train		dev		test	
	sent.	acr.	sent.	acr.	sent.	acr.
English/l	2949	242	385	31	383	30
English/s	7532	405	894	52	574	40
French	7851	541	909	68	813	60
Spanish	6267	437	818	56	862	53

Table 1
Statistical information of datasets

only slows down the training process but also makes it impossible to fix batch size to a constant.

To overcome this obstacle, we decide to pad all dictionaries to a fixed size $k = \max_i\{|D_i|\}$, which makes it convenient for training in a constant batch size. Considering the acronyms in the train/dev/test datasets are exclusive, which means an acronym appearing in the test dataset doesn't show up in the train or dev dataset, we sample pseudo expansion candidates to pad the dictionaries. We call this method **Negative Expansion Sampling**. We also tried to use a meaningless symbol as [PAD] token and mask out their scores before softmax, but it leaded to a worse performance. The negative expansion sampling strategy not only solves the engineering technical problems, but also can strengthen the robustness of our model.

3.3.2. CLIP-Adapter.

Large-scale pre-trained models have shown significant progress in lots of domains, but how to effectively transfer the learned knowledge to downstream tasks remains an open problem. Unlike prompt-based methods, CLIP-Adapter adopts two linear layers to transform the features v extracted by PLMs into new task-specific features v_s and blend them together in a residual-style, which is represented as $v^\star$.

$$v_s = \text{ReLU}\left(v^T \mathbf{W}_1\right)\mathbf{W}_2 \tag{2}$$

$$v^\star = \lambda \cdot v_s + (1-\lambda)\cdot v \tag{3}$$

Experiments show CLIP-Adapter can bridge the semantic gap between the downstream task and the pre-training dataset. In residual blending, a hyperparameter λ is used for trading off original and task-specific information for better performance.

3.3.3. Adversarial Training.

Adversarial training is a widely used technique that can enhances robustness and generalization of the model. The optimization objective can be expressed as follow:

$$\min[\text{CE}(\mathbf{y}, \hat{\mathbf{y}}) + \beta \underbrace{\max_{\|\epsilon\|\leq c} \text{KL}(\hat{\mathbf{y}}, f(\mathbf{x}+\epsilon))}_{R}] \tag{4}$$

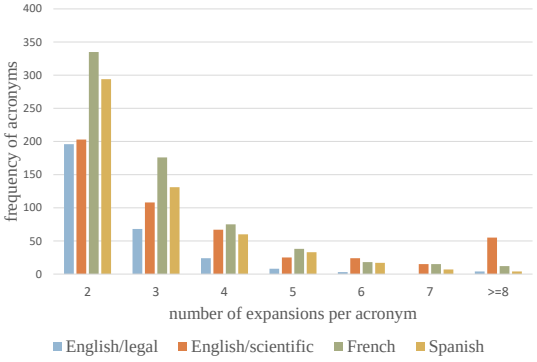

Figure 3: Distribution of expansions per acronym

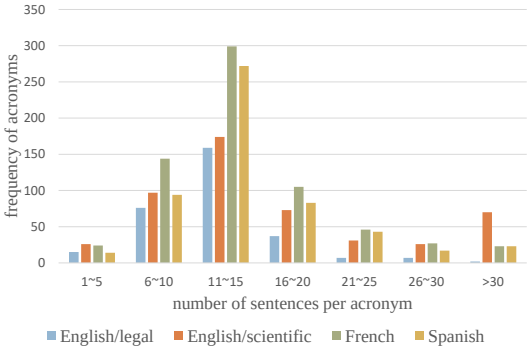

Figure 4: Distribution of sentences per acronym

where CE is the cross entropy loss and KL is Kullback-Leibler divergence, and $\hat{\mathbf{y}}$ is model prediction. Adversarial training essentially minimizes the KL-based worst-case posterior difference between the clean and corrupted inputs, ϵ is the most adversarial direction searched with an inner loop.

3.3.4. Child Tuning.

Fine-tuning a large pre-trained model on downstream tasks is popular nowadays, but fine-tuning directly on limited training data may cause many problems like catastrophic forgetting. Xu et al. [24] proposes a new fine-tune method called Child-Tuning. Specifically, it randomly or strategically masks out part of the gradients during the backward process. Thus only a subset of parameters of pre-trained models is updated. Empirical results show Child-Tuning can help the model obtain better generalization performance.

PLM	English/scientific			English/legal		
	P	R	F1	P	R	F1
BERT$_{large}$	–	–	–	0.708	0.615	0.658
LEGAL-BERT	–	–	–	0.781	0.628	0.696
SciBERT	0.830	0.704	0.762	0.712	0.624	0.665
RoBERTa$_{base}$	0.811	0.765	**0.787**	0.706	0.597	0.647
RoBERTa$_{large}$	–	–	–	0.790	0.637	**0.705**
	French			**Spanish**		
xlm-RoBERTa$_{base}$	0.731	0.654	0.690	0.753	0.683	0.716
BERT-Spanish	–	–	–	0.832	0.795	**0.813**
BERT-multilingual	0.758	0.696	**0.726**	0.810	0.737	0.772

Table 2

Results on test set using different PLMs.

4. Experiments

4.1. Datasets

SDU@AAAI-22 shared task 2 provided multilingual datasets in English(including legal and scientific domain), French, and Spanish[1]. Table 1 shows the statistics of the datasets, where **sent.** and **acr.** represent the number of sentences and acronyms, respectively. And we exclude acronyms that don't appear in any datasets. Besides, **English/l** and **English/s** stand for English in legal and scientific domain. As a zero-shot task, acronyms have no overlap between datasets.

Figure 3 and Figure 4 demonstrate statistical analysis of the datasets. On average, each acronym has 3.1 long-form alternatives, but Figure 3 reveals that the number of expansions per acronym has an apparent long-tail distribution. Meanwhile, each acronym averagely has 15.0 examples, and Figure 4 indicates most acronyms show up in less than 15 sentences.

4.2. Implementation Details

Considering the gap between different languages, we leverage different PLMs trained on the corresponding corpus for each dataset. We utilize RoBERTa [8], BERT-Spanish [19] and BERT-multilingual [6] on English, Spanish and French datasets respectively. What's more, we compare these models to other domain-specific PLMs like SciBERT [20] and LEGAL-BERT [21].

We implement the proposed model based on Hugging Face transformers[25]. The batch size used in our experiments is fixed to 1, because as described above, one sample contains $k = max\{|D_i|\}$ sentences and batch size bigger than 1 may lead to "*out of memory*" error. Though we can increase the batch size by reducing padding size or truncating redundant expansions, we leave that for future research. The learning rate in experiments is 1×10^{-5} and the max epoch is set to 30 with patience equal to 5. We set $\lambda = 0.5$ for all CLIP-Adapter models and add a Dropout layer to it with $p = 0.5$. We perform an adversarial attack on the first layer of the model, i.e., the embedding layer. We use AdamW [26] as our optimizer with a warm-up rate equal to 0.1 and gradient clip norm set to 1.0 in all experiments. We evaluate our model on the development set at the end of each epoch. We use 4 NVIDIA RTX 3090 GPUs for training. Detailed experiment results are described in the next section.

[1]https://github.com/amirveyseh/AAAI-22-SDU-shared-task-2-AD

4.3. Performance Comparison

Table 2 demonstrates the main results of our proposed method using different PLMs as the backbone, in which we can find that the choice of PLMs has an important impact on the performance.

On English datasets, RoBERTa achieves the best score among selected PLMs, which is **0.79** by RoBERTa$_{base}$ on scientific domain and **0.705** by RoBERTa$_{large}$ on legal domain. Compared with BERT$_{large}$, RoBERTa$_{large}$ improves the F1 score by 4.7%, indicating the modified training strategies that RoBERTa adopts have a significant contribution to its performance. Surprisingly, RoBERTa also outperforms domain-specific models like SciBERT and LEGAL-BERT, which implies that more training corpus and larger model size can help PLMs overcome the possible deficiencies in domain-specific knowledge and achieve comparable or even better performance than small-scale domain-specific PLMs.

On French and Spanish datasets, BERT-multilingual and BERT-Spanish get the highest scores, respectively. In this turn, we use xlm-RoBERTa$_{base}$ [27] as the comparative model, which is a multilingual version of RoBERTa. In contrast to English datasets, the degree of language specialization has a crucial impact on French and Spanish.

Model	Precision	Recall	F1
BERT-multilingual	**0.764**	**0.690**	**0.725**
-w/o Negative Expansion Sampling	0.687	0.645	0.665
-w/o CLIP-Adapter	0.719	0.671	0.694
-w/o Child-Tuning	0.710	0.671	0.690
-w/o Adversarial Training	0.731	0.655	0.691

Table 3
Ablation study on the French dev dataset

Model	Precision	Recall	F1
BERT-Spanish	**0.869**	0.813	0.840
-w/o Negative Expansion Sampling	0.811	0.801	0.806
-w/o CLIP-Adapter	0.850	0.805	0.827
-w/o Child-Tuning	0.867	**0.823**	**0.844**
-w/o Adversarial Training	0.847	0.799	0.822

Table 4
Ablation study on the Spanish dev dataset

Especially on Spanish datasets, BERT-Spanish achieves a dominantly superior score of **0.813** and is absolutely 9.7% and 4.1% higher than xlm-RoBERTa$_{base}$ and BERT-multilingual respectively. This huge improvement may be owing to the giant discrepancy between different languages and demonstrates the importance of leveraging language-specific PLMs in cross-linguistic tasks.

4.4. Ablation Study

To evaluate the effectiveness of our training strategies, we conducted an ablation study on both the French and Spanish development dataset, which is shown in Table 3 and 4. As shown in the table, the F1 score of the model without negative expansion sampling decreases by 6% and 3.4% on French and Spanish datasets respectively, which confirms that this strategy has a significant edge on simply padding with meaningless tokens. At the same time, F1 score decreases by 3.1% and 1.3% without CLIP-Adapter, indicating that fusing original and task-specific information can promote the performance. As for Child-Tuning, its effect depends on the datasets, where in French domain F1 score decreases by 3.5% without it, but F1 score increases by 0.4% in Spanish domain. This result may be owing to the sensitivity of large-scale PLM, which can be easily affected by downstream task dataset and should be very carefully tuned. For adversarial training, the absence of this component harms the model performance on both French and Spanish datasets, demonstrating the necessity of using adversarial training.

5. Conclusion

In this paper, we propose a multi-choice classification model for the acronym disambiguation. To solve the problem caused by treating the task as a multi-choice classification problem, we propose Negative Expansion Sampling, which can also strengthen the robustness of our model. We explore various PLMs on the cross-linguistic datasets and utilize several training strategies to further boost the model performance. Comprehensive experiments prove the effectiveness of our proposed model. Moreover, the ablation study confirms the necessity of our training strategies. With all of these indispensable components, our model finally achieves competitive performance on the multilingual acronym disambiguation task.

6. Acknowledgments

This research was supported in part by the National Key Research and Development Program of China (No. 2020YFB1708200) and the Shenzhen Key Laboratory of Marine IntelliSense and Computation under Contract ZDSYS20200811142605016.

References

[1] W. Liu, T. Xu, Q. Xu, J. Song, Y. Zu, An encoding strategy based word-character LSTM for chinese NER, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 2379–2389.

[2] Y. Shi, Y. Yang, Relational facts extraction with splitting mechanism, in: 2020 IEEE International Conference on Knowledge Graph, ICKG 2020„ IEEE, 2020, pp. 374–379.

[3] L. Ding, Z. Lei, G. Xun, Y. Yang, FAT-RE: A faster dependency-free model for relation extraction, J. Web Semant. 65 (2020) 100598.

[4] 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.

[5] 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.

[6] C. Pan, B. Song, S. Wang, Z. Luo, Bert-based acronym disambiguation with multiple training strategies, in: Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021, 2021.

[7] J. Devlin, M. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[9] Q. Zhong, G. Zeng, D. Zhu, Y. Zhang, W. Lin, B. Chen, J. Tang, Leveraging domain agnostic and specific knowledge for acronym disambiguation, in: Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021, 2021.

[10] N. Egan, J. Bohannon, Primer ai's systems for acronym identification and disambiguation, in: Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021, 2021.

[11] A. Singh, P. Kumar, Scidr at SDU-2020 : IDEAS - identifying and disambiguating everyday acronyms for scientific domain, in: Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021, 2021.

[12] J. L. M. Pereira, H. Galhardas, D. E. Shasha, Acronym expander at sdu@aaai-21: an acronym disambiguation module, in: Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021, 2021.

[13] R. Navigli, Word sense disambiguation: A survey, ACM Comput. Surv. (2009) 10:1–10:69.

[14] B. Scarlini, T. Pasini, R. Navigli, Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, 2020, pp. 8758–8765.

[15] M. Bevilacqua, R. Navigli, Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, 2020, pp. 2854–2864.

[16] M. R. Ciosici, T. Sommer, I. Assent, Unsupervised abbreviation disambiguation contextual disambiguation using word embeddings, CoRR abs/1904.00929 (2019).

[17] E. Barba, L. Procopio, N. Campolungo, T. Pasini, R. Navigli, Mulan: Multilingual label propagation for word sense disambiguation, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, 2020, pp. 3837–3844.

[18] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018).

[19] J. Canete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, Pml4dc at iclr (2020) 2020.

[20] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, 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 2019, Hong Kong, China, November 3-7, 2019, 2019, pp. 3613–3618.

[21] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: the muppets straight out of law school, CoRR abs/2010.02559 (2020).

[22] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, Y. Qiao, Clip-adapter: Better vision-language models with feature adapters, CoRR abs/2110.04544 (2021). URL: https://arxiv.org/abs/2110.04544. arXiv:2110.04544.

[23] T. Miyato, S. Maeda, M. Koyama, S. Ishii, Virtual adversarial training: A regularization method for supervised and semi-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. (2019) 1979–1993.

[24] 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, 2021, pp. 9514–9528.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's transformers: State-of-the-art natural language processing, CoRR abs/1910.03771 (2019). URL: http://arxiv.org/abs/1910.03771. arXiv:1910.03771.

[26] I. Loshchilov, F. Hutter, Fixing weight decay regularization in adam, CoRR abs/1711.05101 (2017). URL: http://arxiv.org/abs/1711.05101. arXiv:1711.05101.

[27] 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, ACL 2020, Online, July 5-10, 2020, 2020, pp. 8440–8451.