ANACONDA: Adversarial Training with In-trust Loss in Acronym Disambiguation

Fei Xia (Co-first author)[1,2], Bin Li (Co-first author)[3], Yixuan Weng[1], Xiusheng Huang[1,2] and Shizhu He (Corresponding author)[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
[3]College of Electrical and Information Engineering, Hunan University

Abstract

Acronym Disambiguation (AD) aims to find the correct expansions of an ambiguous acronym in a given sentence, which is essential for scientific document understanding tasks. In supervised AD, a significant challenge is to classify the meaning of most words under low resource conditions. For example, 82.64% of the annotated acronym examples in the legal AD training data are less than 15. This problem becomes more apparent when the distribution of words and senses is unbalanced. In this paper, we propose ANACONDA, an **A**dversarial training framework with i**N**trust loss in **ACrON**ym **D**isambigu**A**tion. Experiments on Legal English show the effectiveness of our proposed methods, and our score ranks 1st in SDU@AAAI-22 shared task 2: Acronym Disambiguation.

Keywords

Acronym Disambiguation, Document Understanding, Adversarial Training

1. Introduction

An acronym is a word created from the initial components of a phrase or name, called the expansion [1, 2]. They are short forms of longer terms, and they are frequently used in writing, especially in scientific documents, to save space and facilitate the communication of information. However, as people increasingly use abbreviations, this introduces more text-understanding challenges, primarily scientific document understanding [3, 4]. More specifically, as the acronyms might not be defined in dictionaries, especially locally-defined acronyms whose long-form is only provided in the document that introduces them, identifying the acronyms and their long-forms correctly in the text is a challenging task.

Acronym disambiguation (AD) aims to determine the correct long form of an ambiguous acronym in a given text [3]. It is usually formulated as a sequence classification problem in general [5]. Figure 1 is an example of this task, given a sentence *"GPS The Mechanism is fundamental to the implementation of the NEPAD priorities of political, economic and corporate governance, a central element in strengthening Africa's ownership of NEPAD and a means of attracting support from development partners"*. In this example, the ambiguous acronym in the input sentence is shown in boldface. The possible expansion (long form) of the acronym will also be given. In this example, *"GPS"* may be 1) *global positioning system*, 2) *Governance, peace and security*, or 3) *Global Positioning System* (uppercase first letter). A sound AD system needs to correctly recognize that the *"GPS"* in the example corresponds to *"Governance, peace and security"*.

Input:
-Sentence: **GPS** The Mechanism is fundamental to the implementation of the NEPAD priorities of political, economic and corporate governance, a central element in strengthening Africa's ownership of NEPAD and a means of attracting support from development partners.

-Dictionary: **GPS:**
1. global positioning system
2. Governance, peace and security
3. Global Positioning System

Output: Governance, peace and security

Figure 1: Example of acronym disambiguation.

In the past few years, thanks to more sophisticated neural methods, the performance of AD tasks has been significantly improved [6]. For example, it combines hand-designed rules [7], hand-made functions [8], word embedding [9] and pre-training techniques [10]. However, due to the lack of high-quality annotation data and the heavy expertise and workload required to expand these materials, the potential of these methods is severely

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

limited. This problem has been affecting many tasks of NLP for a long time, primarily related to word sense disambiguation [11], because the granularity of word meaning is very fine, and it is often difficult to distinguish. If the distribution in the corpus is not balanced, it will further aggravate the difficulty of AD classification.

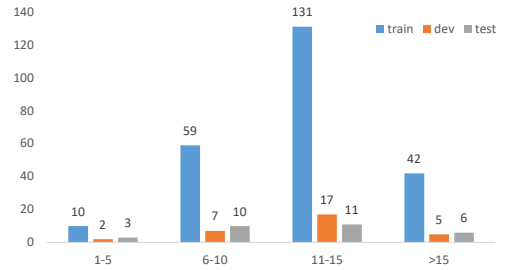

Figure 2: Number of samples per acronym

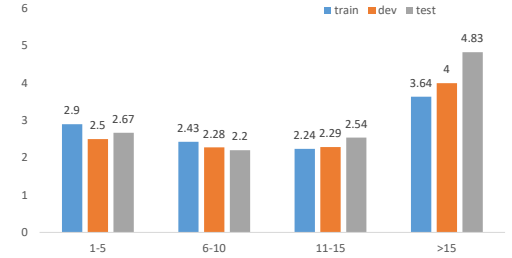

Figure 3: Average number of expansion for acronym under a certain sample frequency.

We have done some analysis on legal English data. As shown in Figure 2, 82.64% of acronyms' samples appear less than 15 times. There are many acronyms with multiple possible extensions, which means that each extension can only have two or fewer examples to learn. It brings difficulties to acronym disambiguation. At the same time, the unbalanced sample distribution of acronyms is also a significant problem. As shown in Figure 3, acronyms with less than 15 samples have an average possible expansion of about 2.5. The acronyms with more than 15 samples have more than four possible expansions on average. If more expansions of acronyms need more samples, then the number of samples between 11-15 should be more than 0-5, but that is not the case. The emergence of this situation also brings difficulties to the acronym disambiguation.

In this paper, we propose ANACONDA, **A**dversarial training with i**N**trust loss in **AC**r**ON**ym **D**isambigu**A**tion. Our purpose is to help the model learn complex samples of acronyms and improve the robustness of the model. Specifically, after analyzing the data, we found that some data have problems such as lack of sufficient labelled samples, complex samples (their meaning very close), and unbalanced data distribution. These problems make it difficult for the model to predict the meaning of acronyms correctly. Therefore, we adopt a dynamic curriculum learning method to dynamically extract complex samples (model predicted error and low-confidence data) from the training data and add them to the training process to let the model learn several times. In addition, we also use adversarial training techniques to improve the robustness of the model. Finally, different from the general cross-entropy loss function, we use the enhanced In-trust loss [12] function to improve the model's generalisation ability further.

The main contributions of this work are summarized as follows:

- We analyze and found several problems that make AD tasks hard to improve, including complex samples, unbalanced data distribution, and provide solutions.
- We propose ANACONDA, **A**dversarial training with i**N**trust loss in **AC**r**ON**ym **D**isambigu**A**tion. This method helps the model learn difficult samples of acronyms and improve the robustness of the model.
- Experiments conducted on the legal English dataset demonstrate that the proposed method has better performance and outperforms other competitive baselines.

2. Related work

As the core task of natural language processing, word sense disambiguation has been extensively studied. Some work has been used to deal with the lack of labelling data. Early work used WordNet's lexical relationships, especially the singular and plural kinship relationships of polysemous words, to calculate correlations [13]. Although these methods prove their ability to generate new training examples, they still need to be improved in cross-language and domain expansion. Later work designed features to build a classifier for specific words [14]. There are also studies to solve these problems using parallel corpora [15] or multilingual knowledge bases [16]. Recently, Stengel-Eskin proposed a neural discriminant architecture for word alignment and applied it to Chinese NER label propagation [17]. As far as we know, this work is the first to use neural word alignment to project word meanings across languages, which is of reference significance. Other work explored more lexical resources, such as knowledge graph structure [18].

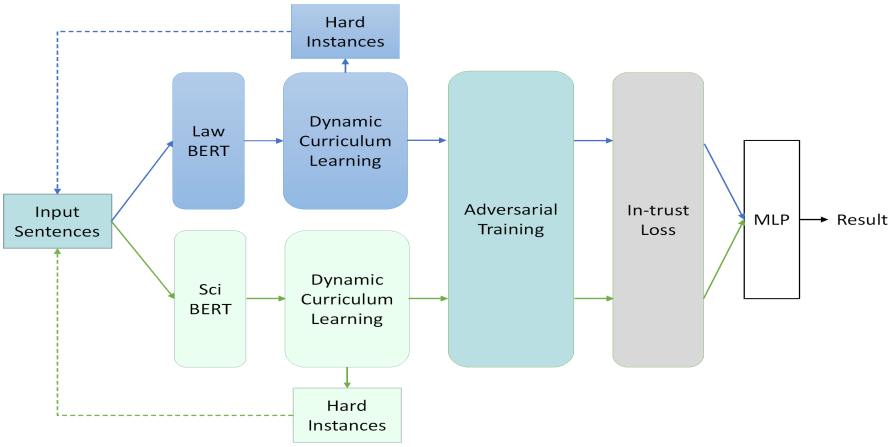

Figure 4: Overview of the proposed method.

3. Task introduction

In this section, we first introduce the problem statement of the acronym disambiguation and then describe the evaluation metric and data.

3.1. Problem Statement

Acronym disambiguation aims to find the correct meaning of an ambiguous acronym in a given sentence. The input $s = w_1, w_2, \ldots, w_n$ is a sentence containing an ambiguous acronym, where n is the total length of the sentence and the acronym is w_i. The dictionary contains all possible extensions $D = d_1, d_2, \ldots, d_k$ corresponding to the acronym,k represents the total length of the probable sentences. The systems are expected to find the correct expanded form d_j of the acronym w_i given the possible expansions D for the acronym.

3.2. Evaluation metric

To evaluate the performance of different methods, the Macro F1 is adopted. The definitions are shown as follows:

$$\text{Precision} = \frac{\sum_{i=1}^{n} \text{precision}_i}{n} \tag{1}$$

$$\text{Recall} = \frac{\sum_{i=1}^{n} \text{recall}_i}{n} \tag{2}$$

$$\text{Macro F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

where n is the number of total classes, the $precision_i$ and $recall_i$ represent the precision and recall of class i respectively.

3.3. Dataset

The data of legal English is shown in Table 1. The data set is divided into training (2949), development (385) and testing (383). The training and validation sets of the legal English data set have been manually labelled, and the labels have been collected into the dictionary.

Table 1
Statistical information of legal English dataset.

Data	Sample Number	Ratio
Training Set	2949	79.33%
Development Set	385	10.36%
Test Set	383	10.30%
Total	3717	100%

4. Method

In this section, we will introduce our overall method framework and present the details of each method separately.

4.1. Model architecture

The overview of our proposed method is shown in the Figure 4. We use LegalBERT [19] in the legal field and SciBERT [20] in the scientific area as two basic models. We first send sentences to the model and use the dynamic curriculum learning [21] method to get hard instances in the input data. Then we send the hard instances to the model for numerous training. At the same time, we also use adversarial training [22], including Fast Gradient Method (FGM) [23] and Projected Gradient Descent (PGD) [24] methods, to increase the learning difficulty of

Table 2
Results in legal English.

Method	Macro Precision	Macro Recall	Macro F1
Rule-based	0.73	0.38	0.5
RoBERTa	0.80	0.72	0.75
SciBERT	0.81	0.73	0.77
LegalBERT	0.83	0.75	0.84
+ Dynamic Curriculum Learning	0.85	0.78	0.86
+ Adversarial Training	0.86	0.80	0.87
+ Enhanced In-trust Loss	0.88	0.83	0.85
Ensemble	**0.94**	**0.87**	**0.90**

simple samples and improve the robustness of the model. In addition, different from the traditional cross-entropy loss function, we use an enhanced In-trust [12] loss function in our task to further improve the model's ability to identify acronyms and expand correctly. Finally, we merge the results obtained by the two models to achieve the best disambiguation effect.

4.2. Dynamic curriculum learning

The main idea of curriculum learning [21] is to imitate the characteristics of human learning. The learning materials of humans and animals are presented in the order of easy to difficult so that the learning effect will be better. Learning the curriculum from simple to complex (in this task are samples that are easy to understand and samples that are not easy to learn), so that it is easy for the model to find a better local optimum, and at the same time speed up the training. Specifically, we send sentences $I = \{s_1, s_2, \dots, s_b\}$ into the models $M_{LegalBERT}$ and $M_{SciBERT}$ and get the prediction result $R = (p, c)$, where b is the size of a batch and p represents whether the prediction is correct, and c represents the prediction confidence. We collect and classify each model's prediction error and low confidence instances S_i, S_j as a hard instance H, and then add it to the training set I again. Through repeated learning, the model will learn the features of complex cases and improve model prediction accuracy. The dynamics are embodied in that as the training deepens, the model will choose differently for hard instances. Therefore, we will dynamically update the set of hard instances in each epoch.

4.3. Adversarial training

In recent years, with the increasing development and implementation of deep learning, adversarial training [22] have also received more and more attention. In NLP, adversarial training is more used as a regularization method to improve the generalization ability[1] of the model.

[1]https://spaces.ac.cn/archives/7234.

The common method in adversarial training is the Fast Gradient Method (FGM) [23]. The idea of FGM is straightforward. Increasing the loss is to increase the gradient so that we can take

$$\Delta x = \epsilon \nabla_x L(x, y; \theta) \tag{4}$$

Where x represents the input, y represents the label, θ is the model parameter, $L(x, y; \theta)$ is the loss of a single sample, Δx is the anti-disturbance.

Of course, to prevent Δx from being too large, it is usually necessary to standardize $\nabla_x L(x, y; \theta)$. The more common way is

$$\Delta x = \epsilon \frac{\nabla_x L(x, y; \theta)}{\|\nabla_x L(x, y; \theta)\|} \tag{5}$$

Another adversarial training method is called Projected Gradient Descent (PGD) [24], which uses multiple iterations to achieve a larger Δx for $L(x + \Delta x, y; \theta)$.

4.4. Enhanced In-trust loss

Traditional classification tasks trust all labelled data, but not all data contribute to models' generalization. Cross-entropy loss is not a good loss function when the data distribution is unbalanced or noisy, especially when the model is over-fitting. Incomplete-Trust (In-trust) [12] loss function, which boosts L_{CRF} with a robust Distrust Cross-Entropy (DCE) term, can effectively alleviate the overfitting caused by previous loss function.

$$L_{DCE} = -p \log(\delta p + (1 - \delta) q) \tag{6}$$

$$L_{\text{In-trust}} = \alpha L_{CE} + \beta L_{DCE} \tag{7}$$

We made further improvements and changed the L_{CRF} to a more appropriate LCE for the previous task, which further improved the effect.

4.5. Experiments

In this section we will introduce baseline models, experimental settings and results.

4.6. Baseline models

- **Rule-based method** The baseline method proposed by Schwartz is a rule-based method [7]. In this baseline, the similarity of the candidate long-forms with the sample text (in terms of several over- lapping words) is first computed. Then, the long-form with the highest similarity score is chosen as the final prediction. The related codes can be found on the website [1].
- **LegalBERT model** The LegalBERT [19] model is a domain-specific pre-trained language model pre-trained on a large number of legal texts. The architecture of LegalBERT follows the same architecture as BERT [25] to capture a well-formed representation of legal data. This model has achieved better performance than the original BERT-based method in some legal tasks and can be regarded as a good backbone for acronym disambiguation.
- **SciBERT model** The SciBERT [20] is a pre-trained language model for science. This architecture of the SciBERT follows the same architecture as BERT [25] to capture the well-formed representation of the scientific data. This model has achieved better performance than the original BERT- based method in some scientific tasks. Law and science have many similarities, so this model is also suitable in the legal field, which can be viewed as a good backbone for the acronym disambiguation.
- **RoBERTa model** The RoBERTa [26] is mainly trained on general domain corpora with Byte Pair Encoding [27] based on the original structure of the BERT. This model can provide a good fine-grained representation of the sentence which can be used in distinguishing acronyms.

4.7. Experimental settings

We conducted experiments on four baseline models, including the rule-based model [7], LegalBERT [19], SciBERT [20] and RoBERTa [26]. All models are implemented based on Huggingface's open-source converter library [28]. We use mixed-precision training [29] based on the Apex library. We use the initial learning rate of 5e-5 for fine-tuning and the AdamW optimizer with a batch size of 32 for optimization. We use the enhanced In-trust loss [12] function to optimize the model.

4.8. Results

Our results on different models and methods are shown in the Table 2. We can find that the rule-based method is far worse than the pre-trained model, and its generalization ability is poor. It can only deal with some predefined acronym ambiguity mechanically. Among the three pre-training models, RoBERTa has the worst effect. LegalBERT has been pre-trained on many texts in the legal field, so it has the best performance and can better identify the ambiguity of acronyms in legal English. Due to the similarity between legal texts and scientific literature, SciBERT, trained on a large amount of scientific documents, performs well. Our experiments show that dynamic curriculum learning, adversarial training and enhanced In-trust loss function methods are effective for this task. Dynamic curriculum learning can help the model learn the features of hard instances. Adversarial training improves the learning difficulty of simple samples and makes the model more robust. The enhanced In-Trust loss function enables the model to learn well even when the data is unbalanced distributed.

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

5. Conclusion

In this paper, we analyze the difficulties of acronym disambiguation in legal English, including hard instances, unbalanced data distribution, and lack of labeled samples. We propose ANACONDA, a framework that combines adversarial training and dynamic curriculum learning with enhanced In-trust loss function. The experimental results respectively reflect the effectiveness of each strategy. Our method achieved the best performance in the acronym disambiguation of legal English, which shows the effectiveness and competitiveness of our methods.

6. 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.

References

[1] K. Jacobs, A. Itai, S. Wintner, Acronyms: identification, expansion and disambiguation, Annals of Mathematics and Artificial Intelligence 88 (2020) 517–532.

[2] N. M. Amir Pouran Ben Veyseh, S. Yoon, R. Jain, F. Dernoncourt, T. H. Nguyen, Multilingual

Acronym Extraction and Disambiguation Shared Tasks at SDU 2022, in: Proceedings of SDU@AAAI-22, 2022.

[3] A. P. B. Veyseh, F. Dernoncourt, T. H. Nguyen, W. Chang, L. A. Celi, Acronym identification and disambiguation shared tasks for scientific document understanding, arXiv preprint arXiv:2012.11760 (2020).

[4] N. M. Amir Pouran Ben Veyseh, S. Yoon, R. Jain, F. Dernoncourt, T. H. Nguyen, MACRONYM: A Large-Scale Dataset for Multilingual and Multi-Domain Acronym Extraction, in: arXiv, 2022.

[5] A. P. B. Veyseh, F. Dernoncourt, Q. H. Tran, T. H. Nguyen, What does this acronym mean? introducing a new dataset for acronym identification and disambiguation, arXiv preprint arXiv:2010.14678 (2020).

[6] M. Bevilacqua, T. Pasini, A. Raganato, R. Navigli, et al., Recent trends in word sense disambiguation: A survey, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conference on Artificial Intelligence, Inc, 2021.

[7] A. S. Schwartz, M. A. Hearst, A simple algorithm for identifying abbreviation definitions in biomedical text, in: Biocomputing 2003, World Scientific, 2002, pp. 451–462.

[8] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, J. Wang, An attention-based bilstm-crf approach to document-level chemical named entity recognition, Bioinformatics 34 (2018) 1381–1388.

[9] A. Jaber, P. Martínez, Participation of uc3m in sdu@ aaai-21: A hybrid approach to disambiguate scientific acronyms., in: SDU@ AAAI, 2021.

[10] Q. Zhong, G. Zeng, D. Zhu, Y. Zhang, W. Lin, B. Chen, J. Tang, Leveraging domain agnostic and specific knowledge for acronym disambiguation., in: SDU@ AAAI, 2021.

[11] B. Scarlini, T. Pasini, R. Navigli, Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 8758–8765.

[12] 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, 2021, pp. 4791–4803.

[13] M. Lesk, Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, in: Proceedings of the 5th annual international conference on Systems documentation, 1986, pp. 24–26.

[14] Z. Zhong, H. T. Ng, It makes sense: A wide-coverage word sense disambiguation system for free text, in: Proceedings of the ACL 2010 system demonstrations, 2010, pp. 78–83.

[15] A. Raganato, C. D. Bovi, R. Navigli, Neural sequence learning models for word sense disambiguation, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp. 1156–1167.

[16] R. Navigli, S. P. Ponzetto, Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial intelligence 193 (2012) 217–250.

[17] E. Stengel-Eskin, T.-r. Su, M. Post, B. Van Durme, A discriminative neural model for cross-lingual word alignment, arXiv preprint arXiv:1909.00444 (2019).

[18] S. Kumar, S. Jat, K. Saxena, P. Talukdar, Zero-shot word sense disambiguation using sense definition embeddings, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5670–5681.

[19] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, arXiv preprint arXiv:2010.02559 (2020).

[20] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).

[21] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th annual international conference on machine learning, 2009, pp. 41–48.

[22] D. Lowd, C. Meek, Adversarial learning, in: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, pp. 641–647.

[23] T. Miyato, A. M. Dai, I. Goodfellow, Adversarial training methods for semi-supervised text classification, arXiv preprint arXiv:1605.07725 (2016).

[24] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).

[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[26] 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, arXiv preprint arXiv:1907.11692 (2019).

[27] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, arXiv preprint arXiv:1508.07909 (2015).

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface's transformers: State-of-

the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).

[29] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, et al., Mixed precision training, arXiv preprint arXiv:1710.03740 (2017).