ADBCMM : Acronym Disambiguation by Building Counterfactuals and Multilingual Mixing

Yixuan Weng[1], **Fei Xia**[1,2], **Bin Li**[3], **Xiusheng Huang**[1,2], **Shizhu He**[1,2], **Kang Liu**[1,2], **Jun Zhao**[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
[1] wengsyx@gmail.com ,{xiafei2020, huangxiusheng2020}@ia.ac.cn, {shizhu.he, kliu, jzhao}@nlpr.ia.ac.cn
[3] libincn@hnu.edu.cn

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, this task is difficult to obtain good performance and receives less attention due to the lack of large amount of annotation data. To address the above issue, this paper proposes an new method for acronym disambiguation, named as ADBCMM, which can significantly improve the performance of low-resource languages by building counterfactuals and multilingual mixing. Specificallyby 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.

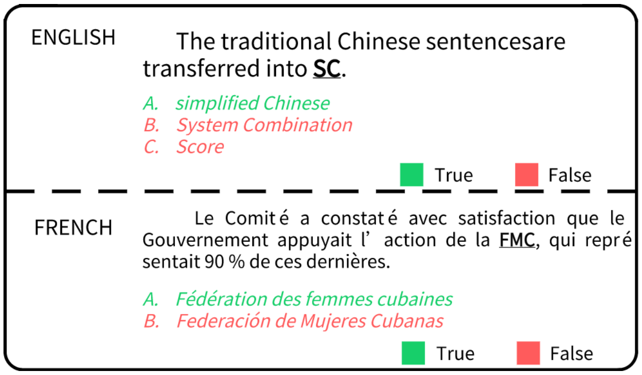
ENGLISH
The traditional Chinese sentencesare transferred into SC.
A. simplified Chinese
B. System Combination
C. Score
🟩 True
🟥 False
FRENCH
Le Comité a constaté avec satisfaction que le Gouvernement appuyait l' action de la FMC, qui représentait 90 % de ces dernières.
A. Fédération des femmes cubaines
B. Federación de Mujeres Cubanas
🟩 True
🟥 False

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

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 when acronyms are used in a large number of scientific documents. 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.

In the datasets, 30,237 data in the four fields of English (science), English (legal), French and Spanish were given. Any data 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 (Li et al. 2018), word embedding (Charbonnier and Wartena 2018), and deep learning (Jin, Liu, and Lu 2019). Over the last few years, the BERT (Devlin et al. 2019) 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 (Pan et al. 2021; Zhong et al. 2021) 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 compar-

isons 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.

Related Work

In this section, we will introduce AD datasets and how to solve AD tasks in English scenarios in the past, while introducing the difficulties of AD tasks in other languages.

AD dataset

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

Table 1: Specific number of AD datasets, including AD tasks for 4 different fields. The total number of data sets is not more than 10,000.

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.

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 (Pan et al. 2021) trained a Binary Classification Model incorporating BERT and several training strategies. His program includes dynamic adverse sample selection, task adaptive pretraining, adversarial training (Goodfellow, Shlens, and Szegedy 2015) and pseudo labelling in his paper. This model achieved its first achievement.

Zhong (Zhong et al. 2021) took into account the field unknowledge and specific knowledge often encountered in AD tasks. 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.

Difficulty in multilingual

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.

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 (Huang et al. 2021), Child-Tuning (Xu et al. 2021) and R-Drop(Liang et al. 2021).

The model framework

We use the Multiple-Choice model framework, which is different from the Binary Classification Model used by Pan (Pan et al. 2021).

The Multiple-Choice model (Wolf et al. 2020) 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 that each option should be by comparing methods. Compared with Binary Classification Model, Multiple-Choice model capturing more semantic characteristics and make the model more comprehensively trained and predicted on differences, rather than the error interference model caused by the dynamic construction of negative samples.

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 (Kaushik et al. 2021). But if man-built counterfactual samples, it would be expensive and time-consuming.

We found many word-like but meaning-different 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

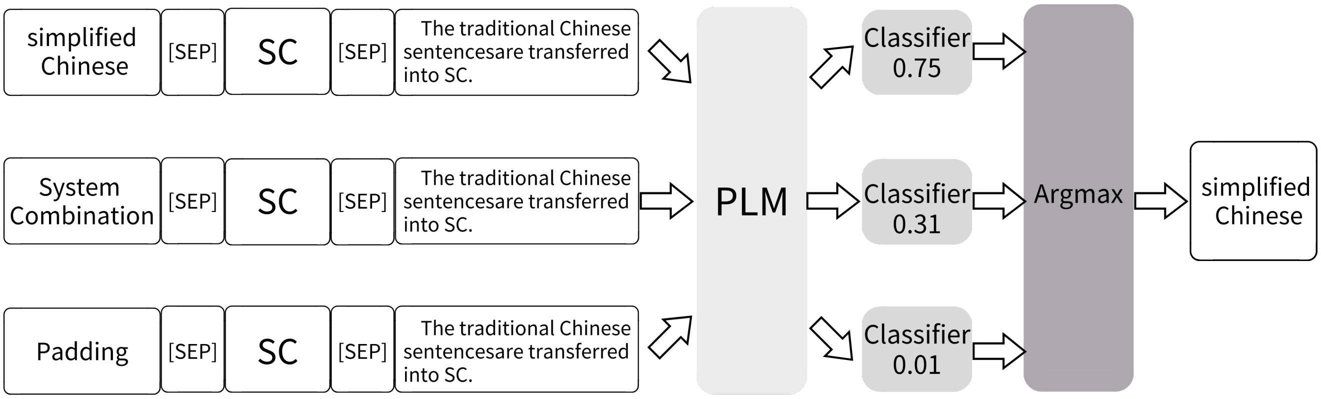
simplified Chinese
[SEP]
SC
[SEP]
The traditional Chinese sentencesare transferred into SC.
System Combination
[SEP]
SC
[SEP]
The traditional Chinese sentencesare transferred into SC.
Padding
[SEP]
SC
[SEP]
The traditional Chinese sentencesare transferred into SC.
PLM
Classifier 0.75
Classifier 0.31
Classifier 0.01
Argmax
simplified Chinese

Figure 2: Multiple-Choice Model

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.

In addition, to address the noise problem of multilingual mixing caused by ADBCMM. We replaced the original CE loss with In-Trust-Loss. This incomplete trust loss function avoids model over-adaptation noise (other languages data) samples while trusting label information and model output. Combined with our ADBCMM method, it has achieved practical results in multilingual hybrid training scenarios.

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.

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 strategy only updates the corresponding Child Network when the parameters are updated backwards, without adjusting all the parameters. This approach like the reverse Dropout (Srivastava et al. 2014), it can bring performance improvements to our models.

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.

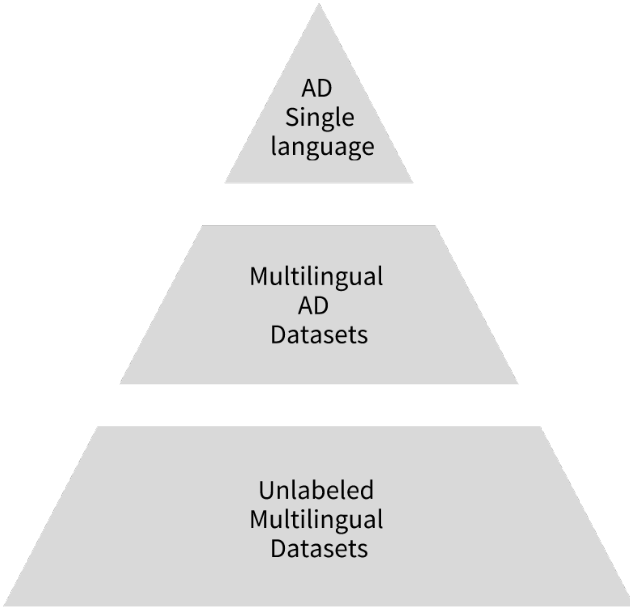

Figure 3: Training Process

Experimental Setting

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

Baseline

For both French and Spanish languages, we used Flaubert-base-cased models and BETO cased models respectively. These models are Bidirectional Encoder Representations from Transformers (Devlin et al. 2019), and the size is both bases. These models have a lot of MLM training in the related large single-language repository and have SOTA re-

	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**

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.

sults 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 AD 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.

Model

To better adapt to the ADBCMM method, we used the DeBERTa model (He, Gao, and Chen 2021) for pre-training in the multilingual repository CC100. The authors of DeBERTa replaced the MLM objective with the RTD (Replaced Token Detection) intent introduced by ELECTRA for pre-training.

Specifically, we used the mdeberta-v3-base 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.

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). We put gradient decrease 1e-5 and batch size 1 (each batch contains 14 different options). We select AdamW Optimizer. We only use the first 300 tokens for each sample. On a 10900K server with 128G memory, we used a 24G NVIDIA 3090 GPU to train our model.

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.[1]

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

[1] 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.

	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

Table 3: SDU@AAAI ranks AD tasks in French and Spanish

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 AD 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 AD tasks, our methods of Precision, Recall and Macro F1 are SOTA. Remarkably, our approach leads us to the second F1 score of 5% - 6%.

Conclusion

In this article, we mainly talk about how to use ADBCMM in AD tasks at SDU@AAAI-22 and compare it with other Models or Methods to ultimately 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.

References

Charbonnier, J.; and Wartena, C. 2018. Using Word Embeddings for Unsupervised Acronym Disambiguation. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2610–2619. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. *CoRR*, abs/1412.6572.

He, P.; Gao, J.; and Chen, W. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv:2111.09543.

Huang, X.; Chen, Y.; Wu, S.; Zhao, J.; Xie, Y.; and Sun, W. 2021. Named Entity Recognition via Noise Aware Training Mechanism with Data Filter. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4791–4803. Online: Association for Computational Linguistics.

Jin, Q.; Liu, J.; and Lu, X. 2019. Deep Contextualized Biomedical Abbreviation Expansion. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, 88–96. Florence, Italy: Association for Computational Linguistics.

Kaushik, D.; Setlur, A.; Hovy, E. H.; and Lipton, Z. C. 2021. Explaining the Efficacy of Counterfactually Augmented Data. In *International Conference on Learning Representations*.

Li, Y.; Zhao, B.; Fuxman, A.; and Tao, F. 2018. 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)*, 1308–1317. Melbourne, Australia: Association for Computational Linguistics.

Liang, X.; Wu, L.; Li, J.; Wang, Y.; Meng, Q.; Qin, T.; Chen, W.; Zhang, M.; and Liu, T.-Y. 2021. R-Drop: Regularized Dropout for Neural Networks. In *NeurIPS*.

Pan, C.; Song, B.; Wang, S.; and Luo, Z. 2021. BERT-based Acronym Disambiguation with Multiple Training Strategies. *ArXiv*, abs/2103.00488.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.

Xu, R.; Luo, F.; Zhang, Z.; Tan, C.; Chang, B.; Huang, S.; and Huang, F. 2021. 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*, 9514–9528. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Zhong, Q.; Zeng, G.; Zhu, D.; Zhang, Y.; Lin, W.; Chen, B.; and Tang, J. 2021. Leveraging Domain Agnostic and Specific Knowledge for Acronym Disambiguation. *ArXiv*, abs/2107.00316.