SimCLAD: A Simple Framework for Contrastive Learning of Acronym Disambiguation

Bin Li (Corresponding author)[1], Fei Xia[2,3], Yixuan Weng[2], Xiusheng Huang[2,3] and Bin Sun[1]

[1]*College of Electrical and Information Engineering, Hunan University*
[2]*National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy Sciences*
[3]*School of Artificial Intelligence, University of Chinese Academy of Sciences*

Abstract

Acronym disambiguation means finding the correct meaning of an ambiguous acronym from the dictionary in a given sentence, which is one of the key points for scientific document understanding (SDU@AAAI-22). Recently, many attempts have tried to solve this problem via fine-tuning the pre-trained masked language models (MLMs) in order to obtain a better acronym representation. However, the acronym meaning is varied under different contexts, whose corresponding phrase representation mapped in different directions lacks discrimination in the entire vector space. Thus, the original representations of the pre-trained MLMs are not ideal for the acronym disambiguation task. In this paper, we propose a **Sim**ple framework for **C**ontrastive **L**earning of **A**cronym **D**isambiguation (**SimCLAD**) method to better understand the acronym meanings. Specifically, we design a continual contrastive pre-training method that enhances the pre-trained model's generalization ability by learning the phrase-level contrastive distributions between true meaning and ambiguous phrases. The results on the acronym disambiguation of the scientific domain in English show that the proposed method outperforms all other competitive state-of-the-art (SOTA) methods.

Keywords

Acronym Disambiguation, Document Understanding, Contrastive Learning, Continual Pte-training

1. Introduction

Recently, the pre-training technology has highly improved the machine understanding level [1]. However, due to the complexity and ambiguity of the natural language, there is still a gap between the machines and humans in comprehensively understanding documents [2]. In scientific document understanding (SDU@AAAI-22), due to space limitations, the appearance of acronyms is often necessary. It is of great significance to correctly understand and distinguish the correct acronym meaning from the given sentence [3].

More precisely, the document reading system is expected to find the correct expanded form of the acronym given the possible expansions from the dictionary for the acronym. This is quite important for a variety of downstream tasks containing the understanding part, such as reading comprehension [4], story cloze [5] and medical entity disambiguation [6], etc.

The acronym disambiguation task aims at finding the correct meaning of the ambiguous acronym in a given text from the dictionary [7]. As shown in Figure 1, the sentence is from the scientific domain in English, where the text in bold represents the short acronym. The dictionary contains the indistinguishable acronym of long-form. Our goal is to predict the correct meaning of the long-form acronyms from the dictionary (i.e., Support vector machines). A good prediction should not only understand the context meaning, but also differ the meaning of ambiguous phrases. Many works have attempted to incorporate the manually designed rules [8], handcrafted features [9, 10], word embedding [11] and pre-training technology [12, 13, 14] into this task and achieved relatively good performance. According to the result of the SDU@AAAI-21 [2], the pre-training method can effectively outperform the rule-based or feature-based method by a large margin. However, the acronym meaning varies in different contexts [15], whose corresponding token representation is anisotropic distribution [16]. For the masked language models (MLMs), the token representation is mapped with a cramped idiomatic distribution [17, 16]. As a result, the MLMs are weak in distinguishing the ambiguous meaning of acronyms, especially in the acronym disambiguation task.

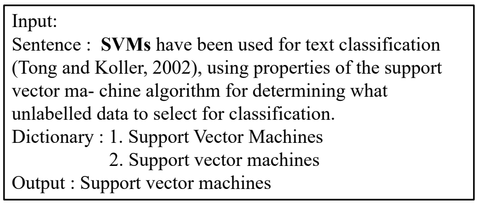
Input:
Sentence : **SVMs** have been used for text classification (Tong and Koller, 2002), using properties of the support vector ma- chine algorithm for determining what unlabelled data to select for classification.
Dictionary : 1. Support Vector Machines
2. Support vector machines
Output : Support vector machines

Figure 1: Example of acronym disambiguation.

SDU@AAAI-22: Workshop on Scientific Document Understanding, co-located with AAAI 2022. 2022 Vancouver, Canada.

libincn@hnu.edu.cn (B. Li (Corresponding author)); xiafei2020@ia.ac.cn (F. Xia); wengsyx@gmail.com (Y. Weng); huangxiusheng2020@ia.ac.cn (X. Huang); sunbin611@hnu.edu.cn (B. Sun)

Inspired by the token-aware contrastive learning method [16], a **Sim**ple framework for **C**ontrastive **L**earning of **A**cronym **D**isambiguation (**Sim-CLAD**) method is proposed to distinguish the distributions between true meaning and ambiguous phrases. Specifically, we adopt the phrase-level continual contrastive pre-training method to enhance the pre-trained MLMs for a better representation of the acronyms. Extensive experiments carried on the acronym disambiguation of the scientific domain in English show that the proposed method achieves the best results compared with the other competitive state-of-the-art (SOTA) methods. The online leaderboard shows that the proposed method ranks 1-st in the scientific English domain in the shared task2 of the SDU@AAAI-22. The main contributions are summarized as follows:

- We perform the first attempt to resolve the acronym disambiguation problem with a contrastive pre-trained model for better acronym understanding.
- We extend the token-level contrastive learning method by designing a phrase-level continuing contrastive pre-training method to obtain better contrastive representations of the ambiguous acronyms.
- Experiments conducted on the scientific English dataset demonstrate that the proposed method has better performance and outperforms other competitive baselines.

2. Related work

2.1. Acronym disambiguation

Acronym disambiguation has attracted much attention in biomedical fields [18]. The earliest methods [8] utilize manually designed rules or text features to find out the acronym expansions. Later, there have been a few works [19] on automatically digging out the acronym expansions by analyzing the web data. These methods are usually effective when an acronym appears in conjunction with the corresponding extensions in the same document. However, traditional rules or statistics cannot effectively handle these tasks with the explosive growth of information. In addition, these methods used for biomedical tasks cannot be directly transferred to other fields, such as science. Recently, deep learning based methods have promoted the development of scientific document understanding (SDU). Methods like feature-based [9], clustering [11], and pre-training model methods [14] perform well in this task. Although these methods based on the pre-training technology (i.e., MLMs) can effectively distinguish confusing phrases of the acronym, they still lack the cognition of negative samples in the representation. Different from the above methods, we use contrast learning to obtain more obvious features for acronym disambiguation.

2.2. Contrastive learning

In general, methods based on contrastive learning (CL) can well distinguish the observed data from other negative samples. Many attempts of the CL have been made to many areas of computer vision, including image [20] and video [21]. Most recently, a simple framework for the CL of visual representations named as SimCLR [22] based on NT-Xent loss is proposed for better image representation. The same idea can also be found in the field of natural language processing (NLP). In the field of NLP, many works [23, 24] are devoted to modeling better sentence-level representations with the CL for the downstream tasks. Recently, Su et al. [16] propose a token-aware CL framework to learn the isotropic and discriminative distribution of token representations by restoring the original token meaning of the masked items. This method is very effective in distinguishing the token-level representations thereby achieving better performance in sentence representation. Following this work, we further consider the phrase-level CL by recovering the probable phrases (i.e., ambiguous acronyms) during the pre-training phase to obtain a better-distinguished acronym representation.

2.3. Continual pre-training

It is a wise choice for further continual pre-training the pre-trained model [25] to alleviate the task and domain discrepancy between the upstream and the downstream tasks. Many works tend to investigate how to better transform the general knowledge to the domain-specific task via continuing pre-training [26, 16]. In the field of the SDU, the generic MLMs are weak in well distinguishing confusing phrases from the dictionary. As a result, the continual pre-training method is adopted in this paper to directly improve the ability of understanding with contrastive learning.

3. Task introduction

3.1. Problem definition

The acronym disambiguation task aims to find the correct meaning of a given acronym in a sentence. Specifically, the sentence can be represented as $X = [x_1, x_2, \ldots, x_n]$, where n is the total length of the sentence. Given that the index i represents the acronym in the input sentence, the short acronym can be represented as $\hat{x}_i$, The corresponding meaning of the short-form acronym is chosen from

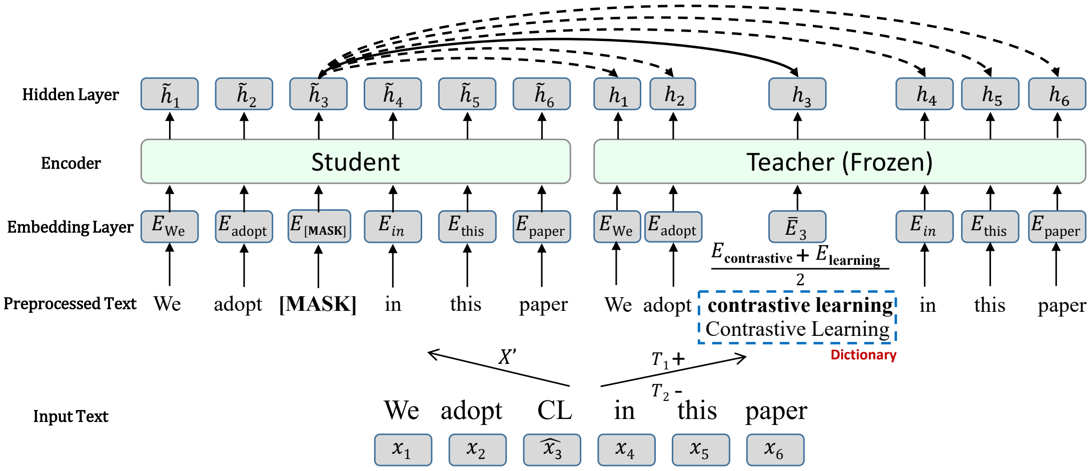
Hidden Layer
Encoder
Embedding Layer
Preprocessed Text
Input Text
$\tilde{h}_1$ $\tilde{h}_2$ $\tilde{h}_3$ $\tilde{h}_4$ $\tilde{h}_5$ $\tilde{h}_6$
h_1 h_2 h_3 h_4 h_5 h_6
Student
Teacher (Frozen)
E_{We} E_{adopt} $E_{\text{[MASK]}}$ E_{in} E_{this} E_{paper}
E_{We} E_{adopt} $\bar{E}_3$ E_{in} E_{this} E_{paper}
$\frac{E_{\text{contrastive}}+E_{\text{learning}}}{2}$
We adopt [MASK] in this paper
We adopt contrastive learning in this paper
Contrastive Learning
Dictionary
X'
T_1+
T_2-
We adopt CL in this paper
x_1 x_2 $\hat{x}_3$ x_4 x_5 x_6

Figure 2: Overview of the proposed method, where the student learns the masked acronym closer to its real meaning produced by the teacher (solid arrow) and away from the other confusing phrases in the dictionary (dashed arrows). The phrase embedding is averaged before encoding.

the dictionary $D = [T_1, T_2, \ldots, T_k]$, where the T_k represents the phrase in the dictionary, and the k represents the total length of the probable phrases. Our goal is to predict the correct phrase meaning S_j of short acronym $\hat{x}_i$ from the dictionary D, where the $j \in [1, k], i \in [1, n]$.

3.2. Evaluation metric

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

$$\begin{aligned} \text{Precision} &= \frac{\sum_{i=1}^{n} \text{precision}_i}{n} \\ \text{Recall} &= \frac{\sum_{i=1}^{n} \text{recall}_i}{n} \\ \text{Macro F1} &= \frac{2 \times \text{ Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \end{aligned} \tag{1}$$

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

Table 1
Statistical information of scientific English dataset.

Data	Sample Number	Ratio
Training Set	7532	83.69%
Development Set	894	9.93%
Test Set	574	6.38%
Total	9000	100%

The acronym disambiguation contains the dataset of scientific English, which is shown in Table 1. The dataset is divided into training (7532), development (894), and testing (574) according to the data set. All the datasets can be found in the work [27], where the training and validation sets of the scientific English dataset have been manually labeled. All the labels are collected in the dictionary.

4. Method

4.1. Model architecture

As shown in Figure 2, the overview of the proposed method contains two domain pre-trained models (a student and a teacher) which are initialized with the same parameters, i.e., SciBERT (Beltagy, Lo, and Cohan 2019). At the stage of pre-training, the parameters of the teacher are frozen to provide a good encoding representation for the student model. In addition, the teacher supports the well-formed original objectives of the MLM (i.e., masked language modeling and next sentence prediction) for the student model. Inspired by [16], we intentionally mask the original short-form acronym ($X^{'}$) to perform the distinguish ambiguous long-form acronyms (T_1+, T_2-) in the teacher model, where notation $+$ and $-$ are positive and negative samples. A contrast loss is adopted in the pre-training process of the student model. Specifically, it is obtained by masking the short-form acronym (i.e., CL) in the input sentence of the student model against the "correct meaning" produced by the teacher without masking the corresponding phrases. To get the representation of the "reference" phrase in the dictionary (dotted frame), we perform phrase averaged method by averaging the embeddings of the tokens (i.e., contrastive learning), which is presented with the upper bar. Meanwhile, we let the representation distance of positive and negative (i.e., Contrastive Learning) samples stay away to enhance the model's ability to distinguish confusing samples.

4.2. Phrase-level contrastive pre-training

The proposed me-thod is composed of two pre-trained models who are both initialized with the SciBERT model, where the one is the student model (noted as S) and the other is the teacher model (noted as T). During the pre-training phase, we only optimize the parameters of S leaving the T model to be frozen. Given an input sentence $X = [x_1, \ldots, x_n]$, we intentionally mask the short acronym $\hat{x}_i$ following the same pre-training task [17]. Then, we feed the masked sentence $X^{'}$ into the student to perform the pre-trained training task. As a result, we obtain the contextual representation $\widetilde{h} = [\widetilde{h}_1, \ldots, \widetilde{h}_n]$ in the student model, where the **[MASK]** is embedded as $E_{\textbf{[mask]}}$. At the same time, the teacher model replaces the corresponding short acronym $\hat{x}_i$ in the original sentence X with the phrase in the dictionary D as input. It is intuitive that the teacher can distinguish all the probable representations with the dictionary, where we want the student model to distinguish the correct phrase meaning through CL. In the end, the well-formed phrase representation is utilized with the averaged embeddings. Thus, the final representation of the recovered sentence $h = [h_1, \ldots, h_n]$ against the corresponding input sentence is produced by the teacher (see Figure 2). Following the work [16], we further refine the proposed phrase-level contrastive pre-training loss

$$\mathcal{L}_{\mathrm{CL}} = -\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}\left(\hat{x}_i, T_k\right) \log \frac{e^{\mathrm{S}\left(\widetilde{h}_i, h_i\right)/\tau}}{\sum_{j=1}^{n} e^{\mathrm{S}\left(\widetilde{h}_i, h_j\right)/\tau}}, \tag{2}$$

where the indicator function $\mathbb{1}(\hat{x}_i, T_k) = 1$ if $\hat{x}_i$ is the masked acronym and short for the corresponding long-form T_k. Otherwise, $\mathbb{1}(\hat{x}_i, T_k) = 0$. We use the τ as the temperature hyper-parameter and the notation $\mathrm{S}(,)$ represents the similarity function, where we choose the cosine function. The K is the number of the all possible long-form acronyms.

4.3. Optimizing objectives

Naturally, the student model lear-ns to distinguish the masked acronym closer to its corresponding "true" representation produced by the teacher and away from the meaning of the other confusing phrases in the sentence. In summary, the acronym representations learned by the student are more discriminative with the confusing phrases, therefore better following an isotropic distribution [16]. Furthermore, the original pre-training method of the MLM [17] is also adopted for learning good document representations, including the masked language modeling task and the next sentence prediction (NSP) task. The overall optimizing objectives are performed as the continual pre-training in the domain-specific corpus, which can be shown as

$$\mathcal{L} = \mathcal{L}_{\mathrm{CL}} + \mathcal{L}_{\mathrm{MLM}} + \mathcal{L}_{\mathrm{NSP}} \tag{3}$$

where the pre-training step is totally unsupervised, which can be carried out with the vast scientific English dataset. Once the pre-trained model is obtained, the student model will be fine-tuned on the acronym disambiguation task.

4.4. Contrastive fine-tuning

Concretely, given the final hidden state, h_x of the input sentence, the representation of the probable phrases can be represented as h_{T_j}. We concatenate the h_x and the h_{T_j} to obtain the feature $\boldsymbol{h}$ for the two classification and contrastive learning, which can be presented as

$$\boldsymbol{h} = \left[h_x; h_{T_j}\right] \tag{4}$$

A non-linear projection layer is added on top of the pre-trained model for obtaining representation. The positive sample is noted as $+$, and the negative sample is noted as $-$. The calculation of two types of the feature can be shown as follows:

$$\begin{aligned} z_i^{+} &= W_2 \,\mathrm{ReLU}\left(W_1 \boldsymbol{h}^{+}\right) \\ z_j^{-} &= W_2 \,\mathrm{ReLU}\left(W_1 \boldsymbol{h}^{-}\right) \end{aligned} \tag{5}$$

Finally, we perform fine-tuning in a multi-task manner and take a weighted average of the two classification losses and the contrastive loss:

$$\mathcal{L} = \frac{(1-\lambda)}{2}\left(\mathcal{L}_{CE}^{+} + \mathcal{L}_{CE}^{-}\right) + \lambda \mathcal{L}_{\mathrm{CL}} \tag{6}$$

where the λ is the weight hyper-parameter.

5. Experiment setup

5.1. Baseline models

- **Rule-based method** The baseline method proposed by Schwartz is a rule-based method [8]. In this baseline, the similarity of the candidate long-forms with the sample text (in terms of several overlapping 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].

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

Table 2
F1 performance in scientific English.

Method	Macro Precision	Macro Recall	Macro F1
Rule-based	0.74	0.37	0.49
RoBERTa	0.81	0.78	0.79
SciBERT	0.85	0.82	0.83
hdBERT	0.89	0.84	0.86
BERT-MT	0.91	0.87	0.89
Our Method	0.94	0.92	0.93
Ensemble	**0.97**	**0.94**	**0.96**

- **RoBERTa model** The RoBERTa [28] is mainly trained on general domain corpora with Byte Pair Encoding[29] 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.
- **SciBERT model** The SciBERT [30] is a domain-specific pre-trained language model for science. This architecture of the SciBERT follows the same architecture as BERT 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, which can be viewed as a good backbone for the acronym disambiguation.
- **hdBERT model** The hdBERT model [12] considers the domain agnostic and specific knowledge, adopting the hierarchical dual-path BERT method jointly trained with fine-grained and high-level specific representations for acronym disambiguation. The context-based pre-trained models including the RoBERTa and the SciBERT are elaborately involved in encoding these two kinds of knowledge respectively. Finally, the multiple layer perception is devised to integrate to output the prediction.
- **BERT-MT model** The BERT-MT method [14] is designed with a binary classification model incorporating the BERT and several training strategies including dynamic negative sample selection, task adaptive pretraining, adversarial training, and pseudo labeling. This method achieves the best performance in the SDU@AAAI-21 competition of the scientific English, which is the strong baseline.

5.2. Pre-training strategies

We use the continuing pre-training strategy with the proposed method using the Sci-BERT model[2]. Except for the dataset of the competition (SDU@AAAI-22), we additionally used the dataset of the SDU@AAAI-21 and the science paper data set[3] for continuing pre-training. The ratio of the positive sample and negative sample keeps 1:2, where the negative sample can be obtained from different short-form acronyms. To obtain good sentence representations, we add the phrase-level contrastive pre-training objective into the pre-training for 200k steps, another 200k is to perform the original BERT pre-training tasks. The training samples are truncated with a maximum length of 300 and the batch size is set as 32. The temperature τ in Eqn. 2 is set to 2e-2. For optimization, we use the same AdamW optimizer [31] with weighted decay. The initial learning rate is 1e-4 for warmups about 10% of the total steps. We implement the pre-training step with 8 NVIDIA 3090 GPUs with 24GB memory.

5.3. Implementation

As for the RoBERTa and the SciBERT model, we fine-tune with initial learning rates of the 5e-5 optimizing via the AdamW optimizer with batch size of 32.

As for the hdBERT and the BERT-MT model, we follow the default setting the same as the paper [14], where the RoBERTa[4] and the SciBERT[5] is adopted from the Transformers of the Huggingface [32].

The BERT-MT model is fine-tuned for 15 epochs with a batch size of 32. The initial learning rate for the encoder is 1e-5, and the others are 5e-4. The minimum learning rate is 5e-7 with the Adam optimizer [33].

Our fine-tuning stage is implemented with a batch size of 32 for 15 epochs. We utilize the trained student model for fine-tuning the test experiments. The AdamW optimizer is adopted with an initial learning rate of 1e-4 and annealed gradually after a warm-up epoch until it reached 1e-5. The weight hyper-parameter λ is set to 0.5 to accelerate the whole training stage.

As for the ensemble part, we fuse the output prob-

[2] https://huggingface.co/allenai/scibert_scivocab_cased

[3] https://huggingface.co/datasets/scientific_papers

[4] https://huggingface.co/roberta-large

[5] https://huggingface.co/allenai/scibert_scivocab_uncased

Table 3
Online leaderboard.

Method	Precision	Recall	F1
Rank1	**0.97**	**0.94**	**0.96**
Rank2	0.95	0.90	0.93
Rank3	0.88	0.82	0.85
Rank4	0.81	0.77	0.79
Rank5	0.81	0.69	0.75
Baseline	0.74	0.37	0.49

ability of the different baselines and add the balanced weights to get the final predictions, where more implemented details can be found in the work [34].

6. Results

The main results of our model and baselines are shown in Table 2. It can be found that the performance of the pre-trained model outperforms the rule-based method since the rule-based method is difficult to pick the correct phrase from confusing acronym options from the dictionary due to its poor generalization. The SciBERT beats the RoBERTa in the three scores, which indicates that the domain-specific pre-training is of significant for science document understanding. The scientific domain pre-trained model can capture a deep representation of the confusing acronyms. The hdBERT merges different types of hidden features to get better generalization in binary classification, thereby performing well in this task. The results of the BERT-MT demonstrate that there are indeed many useful tricks in helping the model enhance the ability of robustness. It is noted that the proposed method outperforms the other baselines in three scores, which represents that the pre-trained model with continuing contrastive pre-training can further improve the model's ability to represent acronyms. Notice that the ensemble method can further improve the diversity of the final results thereby achieving the best performance in the test set. In summary, we finally rank the 1-st in the online leaderboard, which is shown in Table 3.

7. Conclusion

We describe a simple framework for contrastive learning of acronym disambiguation in the shared task 2 of the SDU@AAAI-22. Many baselines are implemented to compare with the proposed method, including methods based on pre-training, combinations of different structures, and useful tricks. The results demonstrate that the proposed method outperforms all other baselines, achieving the best performance (top-1) in the acronym disambiguation of scientific English. It can be further concluded that the continuing contrastive pre-training method can enhance the model's ability to represent the confusing phrases of the long-form acronym. The contrastive fine-tune can further enhance the generalization ability. In future work, we will extend our work as follows: (1) to use twin networks for training the teacher and the student together. (2) Adopting the fine-grained and the coarse-grained embedding into the contrastive pre-training to better acknowledge the meaning of the sentence.

Acknowledgement

This work is supported by the National Key Research and Development Project of China (2018YFB1305200) and the National Natural Science Fund of China (62171183, 61801178).

References

[1] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, Pre-trained models for natural language processing: A survey, Science China Technological Sciences (2020) 1–26.

[2] 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 (2020a).

[3] 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.

[4] M. Gardner, J. Berant, H. Hajishirzi, A. Talmor, S. Min, On making reading comprehension more comprehensive, in: Proceedings of the 2nd Workshop on Machine Reading for Question Answering, 2019, pp. 105–112.

[5] J. Guan, Z. Feng, Y. Chen, R. He, X. Mao, C. Fan, M. Huang, Lot: A benchmark for evaluating chinese long text understanding and generation, arXiv preprint arXiv:2108.12960 (2021).

[6] B. Li, E. Chen, H. Liu, Y. Weng, B. Sun, S. Li, Y. Bai, M. Hu, More but correct: Generating diversified and entity-revised medical response, arXiv preprint arXiv:2108.01266 (2021).

[7] 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, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020b, pp. 3285–3301.

[8] 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.

[9] 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.

[10] F. Li, Z. Mai, W. Zou, W. Ou, X. Qin, Y. Lin, W. Zhang, Systems at sdu-2021 task 1: Transformers for sentence level sequence label., in: SDU@ AAAI, 2021.

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

[12] 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.

[13] D. R. Kubal, A. Nagvenkar, Effective ensembling of transformer based language models for acronyms identification., in: SDU@ AAAI, 2021.

[14] C. Pan, B. Song, S. Wang, Z. Luo, Bert-based acronym disambiguation with multiple training strategies, arXiv preprint arXiv:2103.00488 (2021).

[15] A. P. B. Veyseh, F. Dernoncourt, W. Chang, T. H. Nguyen, Maddog: A web-based system for acronym identification and disambiguation, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2021, pp. 160–167.

[16] Y. Su, F. Liu, Z. Meng, L. Shu, E. Shareghi, N. Collier, Tacl: Improving bert pre-training with token-aware contrastive learning, arXiv preprint arXiv:2111.04198 (2021).

[17] 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), 2019, pp. 4171–4186.

[18] Q. Jin, J. Liu, X. Lu, Deep contextualized biomedical abbreviation expansion, in: Proceedings of the 18th BioNLP Workshop and Shared Task, 2019, pp. 88–96.

[19] D. Nadeau, P. D. Turney, A supervised learning approach to acronym identification, in: Conference of the Canadian Society for Computational Studies of Intelligence, Springer, 2005, pp. 319–329.

[20] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, IEEE, 2005, pp. 539–546.

[21] X. Wang, A. Gupta, Unsupervised learning of visual representations using videos, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2794–2802.

[22] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[23] Z. Wu, S. Wang, J. Gu, M. Khabsa, F. Sun, H. Ma, Clear: Contrastive learning for sentence representation, arXiv preprint arXiv:2012.15466 (2020).

[24] F. Liu, I. Vulić, A. Korhonen, N. Collier, Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Punta Cana, Dominican Republic and Online, 2021. URL: https://arxiv.org/abs/2104.08027.

[25] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8342–8360.

[26] R. Han, X. Ren, N. Peng, Econet: Effective continual pretraining of language models for event temporal reasoning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 5367–5380.

[27] 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.

[28] 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).

[29] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1715–1725.

[30] 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, pp. 3615–3620.

[31] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[32] 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).

[33] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[34] G. P. C. Fung, J. X. Yu, H. Wang, D. W. Cheung, H. Liu, A balanced ensemble approach to weighting classifiers for text classification, in: Sixth International Conference on Data Mining (ICDM'06), IEEE, 2006, pp. 869–873.