Acronym Extraction with Hybrid Strategies

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

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University

Abstract

Acronym extraction plays an important role in scientific document understanding. Recently, the AAAI-22 Workshop on Scientific Document Understanding released multiple high-quality datasets and attracted widespread attention. In this work, we present our hybrid strategies with adversarial training for this task. Specifically, we first apply pre-trained models to obtain contextualized text encoding. Then, on the one hand, we employ a sequence labeling strategy with BiLSTM and CRF to tag each word in a sentence. On the other hand, we use a span selection strategy that directly predicts the acronym and long-form spans. In addition, we adopt adversarial training to further improve the robustness and generalization ability of our models. Experimental results show that both methods outperform strong baselines and rank high on the SDU@AAAI-22 - Shared Task 1: Acronym Extraction, our scores rank 2nd in 4 test sets and 3rd in 3 test sets. Moreover, the ablation study further verifies the effectiveness of each component. Our code is available at https://github.com/carlyoung1999/AAAI-SDU-Task1 .

Keywords

Acronym Extraction, Natural Language Processing, BERT

1. Introduction

An acronym consists of the initial letters of the corresponding terminology and is widely used in scientific documents for its convenience. However, this also makes it difficult to understand scientific documents for both humans and machines. In natural language processing, accurate acronym extraction is beneficial for the downstream applications like question answering [1], definition extraction [2] and relation extraction [3, 4]. Recently, SDU@AAAI-22 released multiple datasets [5] for scientific document understanding, and we focus on the task of acronym extraction [6], which aims to extract acronyms and their corresponding explanations (long-forms); a toy example can be seen in Figure 1.

Traditional approaches utilize rule-based pattern [7] or manual features [8] which are labor-force and time-consumed. Recently, deep learning based methods [9, 10] are preferred for their better performance and end-to-end learning.

In this paper, we propose two strategies for acronym extraction, sequence labeling strategy and span selection strategy. Specifically, we first use pre-trained language models like BERT [11] or RoBERTa [12] to obtain contextualized word representations. Then, we utilize BiLSTM to capture feature interactions between adjacent words further and employ CRF to model the dependency between sequence labels for the sequence labeling strategy. As for the span selection strategy, we use binary taggers to predict the start and end index for acronyms or long-forms. To further improve our models' robustness and generalization ability, we employ adversarial training, which dynamically adds noise to avoid overfitting. These two strategies get comparable performance, and we choose the better one for evaluation according to their performance in the development set. Our contributions are as follows:

- We propose two strategies for acronym extraction, sequence labeling and span selection.
- Our adversarial training further improves the robustness and generalization ability of our models.
- Experiments show that our models outperform strong baselines and rank high in the SDU@AAAI-22 - Shared Task 1: Acronym Extraction.

Input:
Existing methods for learning with noisy labels (LNL) primarily take a loss correction approach.

Output:
Acronym: **LNL**
Long-form: **learning with noisy labels**

Figure 1: An example of Acronym Extraction.

The second workshop on Scientific Document Understanding at AAAI 2022

*Corresponding author.
†These authors contributed equally.
lisiheng21@mails.tsinghua.edu.cn (S. Li);
yangc21@mails.tsinghua.edu.cn (C. Yang);
liangt21@mails.tsinghua.edu.cn (T. Liang);
zhuxy21@mails.tsinghua.edu.cn (X. Zhu);
ycz21@mails.tsinghua.edu.cn (C. Yu);
yang.yujiu@sz.tsinghua.edu.cn (Y. Yang)

2. Related Works

In this section, we introduce the related studies for acronym extraction, including Rule-based, LSTM-based, and Pre-trained-based methods.

2.1. Rule-based

Traditional acronym extraction methods mainly focus on rule-based methods. Specifically, most of them [13] utilize generic rules or text patterns to discover acronym expansions in the field of biomedicine. Torres-Schumann and Schulz [14] further extend rule sets to hidden Markov models and improve both recall and precision values. Recently, a new work [15] has made a comprehensive introduction to the rule-based machine identification methods. They comprehensively classify present Rule-based models, analyze two separate approaches (a machine algorithm and a crowd-sourcing approach), and compare them in detail. However, Due to the conservative nature of rule-based models, this method requires complicated manual formulations and lacks flexibility.

2.2. LSTM-based

Taking advantage of LSTM [16]'s power for text modeling, LSTM-based methods has got decent performance in acronym extraction. They mainly focus on better semantic representations and attention mechanisms. DECBAE [17] extracts contextualized features with BioELMo [18] and provides these features to specific abbreviated BiLSTMs, achieving good performance. In addition, they use a simple but effective heuristic method for automatically collecting datasets from a large corpus. Li et al. [19] propose a novel topic-attention model and compare the performance of different attention mechanisms embedded in LSTM and ELMo. Their model is applied to the acronym task of medical terms. To further capture the dependency between sequence labels, Veyseh et al. [20] propose to combine LSTM with CRF for Acronym identification and Disambiguation.

2.3. Pre-trained-based

Language models pre-trained with a large corpus have shown promising performance in lots of downstream tasks. One of the most popular is Bidirectional Encoder Representations from Transformers (BERT) [21], which obtains rich semantic representations by Masked LM task in the pre-training stage. BERT has been applied to many NLP tasks like information extraction [22] and dialogue state tracking [23].

In addition, it is worth mentioning that there have been many fine-grained improvements or specific domain variants of BERT. RoBERTa [12] optimizes the training strategy with BPE (Byte-Pair-Encoding) and dynamic masking to increase shared vocabulary, thus providing more fine-grained representations and stronger robustness. SciBERT [24] has the same structure as BERT, while it is well pre-trained to process scientific documents specifically. Many works utilize the power of pre-trained models for acronym extraction. Pan et al. [25] proposes a multi-task learning method based on BERT-CRF and BERT-Span, which makes full use of these two separate models through redefining the fusion loss function and achieves great performance. Li et al. [26] utilizes Sentence Piece byte-pair encoding to relabel sentences. Then, they are embedded into the XLNet [27] for processing.

3. Methodology

3.1. Task Formulation

Given a text $X = \{x_1, x_2, ..., x_l\}$ where each x_i is a word and l represents text length, acronym extraction aims to find all acronyms and long-forms mentioned in this text. Formally, the model needs to automatically extract acronym mention set $\mathscr{A} = \{[s_1, e_1), [s_2, e_2), ..., [s_n, e_n)\}$, where s_i and e_i denotes the start and end position of the i-th acronym respectively. In addition, the model also needs to extract long-form mention set $\mathscr{B} = \{[s_1, e_1), [s_2, e_2), ..., [s_m, e_m)\}$, similar with $\mathscr{A}$.

3.2. Overview

We describe our hybrid strategies to extract acronyms and long-forms in this section. At first, we use pre-trained models for tokenizing and encoding the original sentence. Then, we employ a BiLSTM-CRF head to model acronym extraction as a sequence labeling task and a BiLSTM-Span head to model it as a span selection task. In addition, to improve the robustness and generalization of our models, we apply adversarial training techniques.

3.3. BERT Encoder

We adopt BERT or RoBERTa as a text encoder to capture rich contextualized word embeddings. For brevity, we use BERT to indicate both BERT and RoBERTa following. Given the input $X = \{x_1, x_2, ..., x_l\}$, with the help of deep multi-head attention layers, BERT captures contextualized representation for each token. The encoding process is as follows:

$$H = \text{BERT}([x_1, x_2, ..., x_l]) = [h_1, h_2, ..., h_l]^T, \quad (1)$$

where $H \in \mathbb{R}^{l \times d}$, and d denotes hidden dimension.

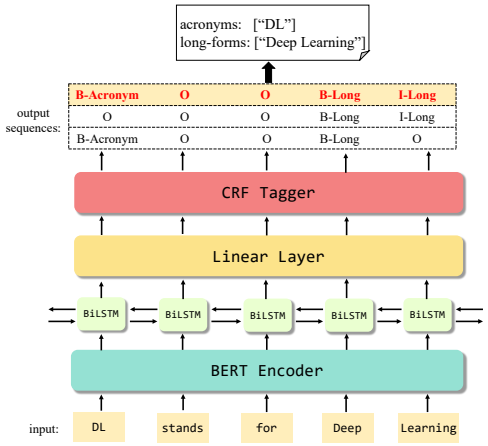

Figure 2: The model architecture of our Sequence Labeling strategy.

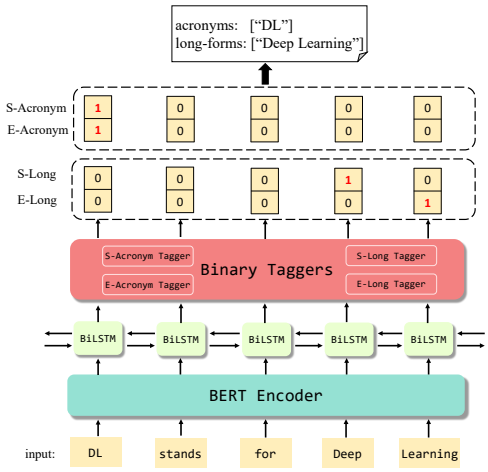

Figure 3: The model architecture of our Span Selection strategy.

3.4. Sequence Labeling Strategy

For this strategy, we first transform the character-level position labels provided by raw datasets to token-level BIO labels as follows:

- **B-Acronym**: Beginning of an acronym.
- **I-Acronym**: Inside of an acronym.
- **B-Long**: Beginning of a long-form.
- **I-Long**: Inside of a long-form.
- **O**: Outside of any acronym and long-form.

To solve this sequence labeling problem, we adopt a BERT-BiLSTM-CRF method, and the architecture is shown in Figure 2. First, we utilize a BiLSTM network to capture feature interactions between adjacent words further:

$$H' = \text{BiLSTM}(H), \tag{2}$$

where $H' \in \mathbb{R}^{l \times 2d}$. Then, a linear classifier transforms H' into the logits of 5 BIO labels defined above:

$$L = [L_0, L_1, L_2, L_3, L_4] = H'W_L, \tag{3}$$

where $W_L \in \mathbb{R}^{2d \times 5}$ and $L = [L_0, L_1, L_2, L_3, L_4] \in \mathbb{R}^{l \times 5}$ are the logits.

To model the dependency between sequence labels, we adopt a Linear Chain CRF (Conditional Random Field) [28], the probability of a tagged sequence is:

$$P(Y|X) = \frac{\exp(\sum_{i=1}^{l} \varphi(y_i|x_i) + \sum_{i=1}^{l} \psi(y_i|y_{i-1}))}{Z(X)}, \tag{4}$$

where $Y = [y_1, y_2, ..., y_l]$ is the ground truth label sequence and y_i is the label for i-th token. $\varphi(\cdot)$ represents emission scorer which refers to the logits L above. $\psi(\cdot)$ denotes transition scorer in CRF and is a learnable matrix practically. $Z(X)$ is the normalization factor to constraint the probability in $(0, 1)$. The loss function is negative log-likelihood :

$$\mathscr{L}_{SL} = -\log(P(Y|X)). \tag{5}$$

For the inference, we use the Viterbi algorithm [28] for decoding the best label sequence.

3.5. Span Selection Strategy

We also formulate it as an extractive span selection task, aiming to find the text span of acronyms and long-forms directly. Similar to the sequence labeling strategy, we transform the character-level labels $[start, end)$ provided by raw datasets to token-level $[start, end)$ for the following token classification.

We adopt the same BERT encoder and LSTM network as above to get contextualized word representations $H' \in \mathbb{R}^{l \times 2d}$. Then we construct four binary taggers:

- **S-Acronym Tagger** predicts whether a token is the start of an acronym.
- **E-Acronym Tagger** predicts whether a token is the end of an acronym.
- **S-Long Tagger** predicts whether a token is the start of a long-form.
- **E-long Tagger** predicts whether a token is the end of a long-form.

We apply a simple linear layer to represent these taggers which work as follows:

$$L = [L_0, L_1, L_2, L_3] = H'W_S, \tag{6}$$

Method	English			Persian			Vietnamese		
	P	R	F1	P	R	F1	P	R	F1
Rule	0.33	0.15	0.20	**0.95**	0.44	0.60	0.82	0.39	0.53
BERT	0.82	0.85	0.83	0.94	0.47	0.63	0.82	0.73	0.77
RoBERTa	0.84	0.88	0.86	0.94	0.52	0.67	0.97	0.48	0.64
Ours-SL	**0.86**	0.88	**0.87**	0.87	0.59	0.70	**0.98**	0.65	0.78
Ours-SS	**0.86**	**0.89**	**0.87**	0.84	**0.67**	**0.73**	0.81	**0.91**	**0.85**

Table 1
Performance comparison on the development sets of scientific domain.

Ranking	English			Persian			Vietnamese		
	P	R	F1	P	R	F1	P	R	F1
1	**0.89**	**0.92**	**0.90**	0.76	**0.82**	**0.79**	0.85	0.82	**0.84**
2	**0.89***	0.89*	0.89*	0.60*	0.69*	0.63*	0.83	**0.84**	0.83
3	0.85	0.87	0.86	**0.92**	0.43	0.59	**0.96***	0.62*	0.76*
4	0.83	0.88	0.86	0.64	0.51	0.57	0.64	0.66	0.65

Table 2
Performance comparison on the test sets of scientific domain, $\star$ indicates the score of our model.

where $W_S \in \mathbb{R}^{2d\times 4}$, and $L = [L_0, L_1, L_2, L_3] \in \mathbb{R}^{l\times 4}$ are logits for 4 classes declared above. The loss function is binary cross entropy:

$$\mathscr{L}_{SS} = \sum_{i=0}^{l}\sum_{j=0}^{3}[-y_i^j\cdot\log(\sigma(l_i^j))+(1-y_i^j)\cdot\log(1-\sigma(l_i^j))], \quad (7)$$

where y_i^j is the label for i-th token regarding class j, l_i^j is the logit for i-th token regarding class j, and $\sigma(x)$ denotes sigmoid function.

For the inference, we first predict the class label of each token. Then, we match each S-Acronym token with the nearest E-Acronym token to get an acronym. The operation for long-form is the same.

3.6. Adversarial Training

To enhance the robustness and generalization ability of our models, we adopt adversarial training. Specifically, given an input X, we incorporate a posterior regularization mechanism [29]:

$$\mathscr{L}_{Adv} = \max_{\|\epsilon\|\leq a}\sum \mathrm{Div}(f_\theta(X)\|f_\theta(X+\epsilon)), \quad (8)$$

where Div is some f-divergence[1], ϵ is noise, a is noise norm and f_θ represents the predict function in our models, like CRF tagger and Binary taggers. This loss regularizes the posterior difference between original and noisy inputs to avoid overfitting. Practically, we use an inner loop to search the most adversarial direction.

[1]We use Jensen-Shannon divergence in our experiments.

Datasets	Training	Development	Test
English Scientific	3980	497	498
Persian	1336	167	168
Vietnamese	1274	159	160
English Legal	3564	445	446
French	7783	973	973
Spanish	5928	741	741
Danish	3082	385	386

Table 3
Statistics of the datasets, the first three belongs to scientific domain while the others belongs to legal domain.

3.7. Objective Function

We jointly train our models with adversarial training, for sequence labeling strategy:

$$\mathscr{L} = \mathscr{L}_{SL} + \alpha\mathscr{L}_{Adv}. \quad (9)$$

For span selection strategy:

$$\mathscr{L} = \mathscr{L}_{SS} + \alpha\mathscr{L}_{Adv}. \quad (10)$$

The α is used for controlling the significance of adversarial training.

4. Experiments

4.1. Datasets

Our experiments are conducted on the official dataset of SDU@AAAI-22 - Shared Task 1: Acronym Extrac-

Method	English			French			Spanish			Danish		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Rule	0.32	0.10	0.16	0.22	0.06	0.10	0.17	0.07	0.10	0.10	0.06	0.08
BERT	0.88	0.87	0.88	0.94	**0.94**	**0.94**	0.89	**0.90**	0.89	0.93	0.94	0.93
RoBERTa	0.87	0.88	0.88	0.78	0.76	0.77	0.88	0.88	0.88	0.90	0.92	0.91
Ours-SL	0.88	0.88	0.88	**0.95**	**0.94**	**0.94**	**0.90**	**0.90**	**0.90**	0.94	**0.95**	**0.94**
Ours-SS	**0.89**	**0.88**	**0.89**	**0.95**	0.93	**0.94**	**0.90**	**0.90**	**0.90**	**0.95**	0.93	**0.94**

Table 4
Performance comparison on the development sets of legal domain.

Ranking	English			French			Spanish			Danish		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
1	**0.90**	**0.92**	**0.91**	**0.94**	**0.95**	**0.94**	**0.90**	**0.91**	**0.91**	**0.95**	**0.98**	**0.96**
2	0.88⋆	0.91⋆	0.90⋆	0.92⋆	0.93⋆	0.93⋆	**0.90**	**0.91**	0.90	**0.95**	0.97	**0.96**
3	0.87	0.91	0.89	0.93	0.92	0.92	**0.90**⋆	0.90⋆	0.90⋆	**0.95**⋆	0.95⋆	0.95⋆
4	0.87	0.90	0.88	0.81	0.80	0.81	**0.90**	0.90	0.90	0.89	0.90	0.89

Table 5
Performance comparison on the test sets of legal domain, ⋆ indicates the score of our model.

Method	English Scientific		
	P	R	F1
Ours-SL	**0.86**	**0.88**	**0.87**
Ours-SL w/o CRF	0.84	**0.88**	0.86
Ours-SL w/o AT	**0.86**	0.87	0.86
Ours-SS	**0.86**	**0.89**	**0.87**
Ours-SS w/o AT	**0.86**	0.87	0.86

Table 6
Ablation studies in the development set of English Scientific.

tion. They provide the data of scientific domain including English, Persian and Vietnamese; and legal domain including English, French, Spanish and Danish. Table 3 summarizes the statistics of datasets used in our experiments.

4.2. Baselines

To investigate the effectiveness of our proposed approach, we compare it with the following three baselines:

- **Rule-based** This method utilizes a manually designed pattern to extract acronyms and is provided by SDU@AAAI-22 [2].
- **BERT-based** This method employs BERT [21] as a text encoder to get contextualized word representation, then employs a classification head to tag each word.
- **Roberta-based** This is similar with above, except RoBERTa [12] as text encoder.

4.3. Implementations

For baselines, we select pre-train models trained with corresponding language corpora in HuggingFace Transformers [30]. As for ours, we adopt the best pre-trained models according to their performance in the development set. Specifically, we adopt roberta-base[3] for English, roberta-fa-zwnj-base-ner[4] for Persian, bert-base-vi-cased[5] for Vietnamese, bert-base-fr-cased[6] for French, bert-base-es-cased[7] for Spanish and danish-bert-botxo-ner-dane[8] for Danish.

We tune the hyper-parameters according to the performance in the development set. For the sequence labeling strategy, the batch size, LSTM layer, LSTM hidden size, adversarial training weight are 8, 1, 256, 0.1, respectively. The batch size, LSTM layer, LSTM hidden size, and adversarial training weight for our span selection strategy are 16, 1, 256, and 1.0. We run all experiments using PyTorch 1.9.1 on the Nvidia GeForce RTX 3090 GPU, Intel(R) Xeon(R) Platinum 8260L CPU on Ubuntu 18.04.4 LTS OS. Our code will be released soon.

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

[3]https://huggingface.co/roberta-base
[4]https://huggingface.co/HooshvareLab/roberta-fa-zwnj-base-ner
[5]https://huggingface.co/Geotrend/bert-base-vi-cased
[6]https://huggingface.co/Geotrend/bert-base-fr-cased
[7]https://huggingface.co/Geotrend/bert-base-es-cased
[8]https://huggingface.co/Maltehb/danish-bert-botxo-ner-dane

4.4. Results

4.4.1. Scientific Domain

The comparison between the proposed model and baseline models is shown in Table 1. The main observations can be summarized as follows:

- Compared with manually designed rule-based methods, pre-trained model-based methods have huge advantages because they can capture reasonable word representations.
- The difference between the BERT model and RoBERTa model is remarkable. We conjecture this is due to the datasets being small; thus, the results depend more on the power of the pre-trained model.
- Our two strategies get similar results and outperform all baseline methods. We submit the better one for testing.

Table 2 shows the top 4 scores in the test sets of the scientific domain; our method gets decent performance and ranks 2st in English and Persian, 3st in Vietnamese.

4.4.2. Legal Domain

The comparison is shown in Table 4, the observations are similar with Scientific Domain, and our method outperforms all baseline models stably. Table 5 shows the top 4 scores in the test sets; our method gets decent performance and ranks 2 in English and French, 3 in Spanish and Danish.

4.5. Ablation Study

To further prove the effectiveness of each component, we run ablation studies on the development set of English Scientific, as shown in Table 6. We find that: (1) for our sequence labeling strategy, CRF is necessary because it helps capture the dependency between sequence labels. (2) adversarial training is beneficial to both strategies by adding reasonable noises, which improve our models' robustness and generalization performance.

5. Conclusion

In this paper, we explore and propose two strategies with adversarial training for SDU@AAAI-22 - Shared Task 1: Acronym Extraction. Experiments show that our methods outperform strong baseline methods in all 7 datasets. In addition, our score ranks high in the test sets. For future work, we will try to solve the problem of class imbalance in both strategies.

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] C. F. Ackermann, C. E. Beller, S. A. Boxwell, E. G. Katz, K. M. Summers, Resolution of acronyms in question answering systems, 2020. US Patent 10,572,597.

[2] D. Kang, A. Head, R. Sidhu, K. Lo, D. S. Weld, M. A. Hearst, Document-level definition detection in scholarly documents: Existing models, error analyses, and future directions, in: Proceedings of SDP@EMNLP 2020, Association for Computational Linguistics, 2020, pp. 196–206.

[3] 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.

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

[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] 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.

[7] N. Okazaki, S. Ananiadou, Building an abbreviation dictionary using a term recognition approach, Bioinform. 22 (2006) 3089–3095.

[8] C. Kuo, M. H. T. Ling, K. Lin, C. Hsu, BIOADI: a machine learning approach to identifying abbreviations and definitions in biological literature, BMC Bioinform. 10 (2009) 7.

[9] D. Zhu, W. Lin, Y. Zhang, Q. Zhong, G. Zeng, W. Wu, J. Tang, AT-BERT: adversarial training BERT for acronym identification winning solution for sdu@aaai-21, CoRR abs/2101.03700 (2021).

[10] N. Egan, J. Bohannon, Primer ai's systems for acronym identification and disambiguation, in: Proceedings of the SDU@AAAI 2021, volume 2831 of *CEUR Workshop Proceedings*, 2021.

[11] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the NAACL-HLT 2019, Volume 1 (Long and Short Pa-

pers), Association for Computational Linguistics, 2019, pp. 4171–4186.

[12] 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, CoRR abs/1907.11692 (2019).

[13] A. S. Schwartz, M. A. Hearst, A simple algorithm for identifying abbreviation definitions in biomedical text, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 4 (2003) 451–462.

[14] E. Torres-Schumann, K. U. Schulz, Stable methods for recognizing acronym-expansion pairs: from rule sets to hidden markov models, Int. J. Document Anal. Recognit. 8 (2006).

[15] C. G. Harris, P. Srinivasan, My word! machine versus human computation methods for identifying and resolving acronyms, Computación y Sistemas 23 (2019).

[16] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.

[17] Q. Jin, J. Liu, X. Lu, Deep contextualized biomedical abbreviation expansion, in: Proceedings of the BioNLP@ACL 2019, Association for Computational Linguistics, 2019, pp. 88–96.

[18] Q. Jin, B. Dhingra, W. W. Cohen, X. Lu, Probing biomedical embeddings from language models, CoRR abs/1904.02181 (2019).

[19] I. Li, M. Yasunaga, M. Y. Nuzumlali, C. Caraballo, S. Mahajan, H. M. Krumholz, D. R. Radev, A neural topic-attention model for medical term abbreviation disambiguation, CoRR abs/1910.14076 (2019).

[20] 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).

[21] 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).

[22] Z. Wei, J. Su, Y. Wang, Y. Tian, Y. Chang, A novel cascade binary tagging framework for relational triple extraction, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Association for Computational Linguistics, 2020, pp. 1476–1488.

[23] S. Kim, S. Yang, G. Kim, S. Lee, Efficient dialogue state tracking by selectively overwriting memory, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Association for Computational Linguistics, 2020, pp. 567–582.

[24] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: Proceedings of the EMNLP-IJCNLP 2019, Association for Computational Linguistics, 2019, pp. 3613–3618.

[25] C. Pan, B. Song, S. Wang, Z. Luo, Bert-based acronym disambiguation with multiple training strategies, in: Proceedings of the SDU@AAAI 2021, volume 2831 of *CEUR Workshop Proceedings*, 2021.

[26] 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: Proceedings of the SDU@AAAI 2021, volume 2831 of *CEUR Workshop Proceedings*, 2021.

[27] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: Advances in Neural Information Processing Systems 32, NeurIPS 2019, 2019, pp. 5754–5764.

[28] C. Sutton, A. McCallum, An introduction to conditional random fields, Found. Trends Mach. Learn. 4 (2012) 267–373.

[29] H. Cheng, X. Liu, L. Pereira, Y. Yu, J. Gao, Posterior differential regularization with f-divergence for improving model robustness, in: Proceedings of the NAACL-HLT 2021, Association for Computational Linguistics, 2021, pp. 1078–1089.

[30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the EMNLP 2020 - Demos, Association for Computational Linguistics, Online, 2020, pp. 38–45.