A Novel Initial Reminder Framework for Acronym Extraction

Xiusheng Huang[1,2], Bin Li[3], Fei Xia[1,2] and Yixuan Weng[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 extraction is committed to extracting acronyms (e.g., short-forms) and their meaning (e.g., long-forms) from the original document, this is one of the key and challenging tasks in scientific document understanding (SDU@AAAI-22) tasks. Previous work regarded them as a task of named entity recognition, ignoring the relationship between acronyms and their meaning, especially the importance of initials. In this paper, we propose a novel Initial Reminder Framework (IRF) for acronym extraction task. Specifically, the IRF recognize the span of acronym for the first time, combined with the initial information, and recognized their meaning again. At the same time, considering that acronyms are often close to their meaning, the IRF adopts Neighborhood Search Strategy. Experiments on two acronym extraction dataset show IRF outperforms the previous methods by 5.90/7.10 F1. Further analysis reveals IRF is effective in extracting short-forms and long-forms.

Keywords

Acronym extractions, The initials, Initial Reminder Framework, Neighborhood Search Strategy

1. Introduction

Acronym extraction is a task to identify acronyms and their meanings, which is very important for scientific document understanding (SDU@AAAI-22)[1, 2]. The previous method regards this task more as a sequence annotation task[3, 4, 5], and the model will recognize the acronyms and long-term.

The context of acronyms often have more obvious characteristics, for example, there are brackets around acronyms, or acronyms themselves have a specific format, which leads to a higher accuracy of identifying acronyms. However, the accuracy of identifying long-term is relatively low, and there are some problems, such as inaccurate identification and no identification.

As shown in Figure 1, in a document, we need to identify the acronyms and long-term. The context of acronyms often has some characteristics (e.g. brackets), which helps the model to identify them. Long-term recognition is a challenge. It needs to have a certain understanding of the document content. The better solution is to know what the corresponding acronym is before extracting long term, which will help model recognition of long term.

Through Figure 1, we can find that each character of the acronym can correspond to the initial of the long-term, which will help the model identify the long-term.

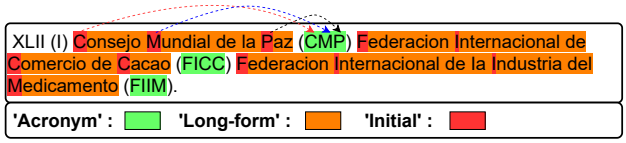

Figure 1: In the figure, the Green text represents acronyms, orange text represents long term, and red text represents initials. At the same time, red, blue and black lines indicate the correspondence between initials and acronyms, respectively. (Dataset: Spanish)

In this paper, we propose a novel Initial Reminder Framework (IRF) for acronym extraction task. Through experiments, we find that the model has high accuracy in acronym recognition than long-term recognition. Specifically, in Spanish, the model achieved 91% F1 in the task of identifying acronyms, the F1 score is only 83%. At the same time, considering the correlation between acronyms and long-term, IRF first completes the task of identifying acronyms. On this basis, combined with the initial information contained in acronyms, IRF further identifies long-term. We verify the effectiveness of our method on two acronym extraction data sets, including Spanish and Danish.

We summarize our contributions as follows:

- We introduce a fresh perspective to revisit the acronym extraction task with a principled problem formulation, which implies a general algorithmic framework that helps the identify long-term by initials.

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

- we propose a novel Initial Reminder Framework (IRF) for acronym extraction task. Specifically, IRF makes use of the high accuracy of acronym recognition and helps the model recognize long-term by integrating the initial information.
- We conduct experiments on two acronym extraction datasets. Experimental results demonstrate that our IRF model can achieves state-of-the-art performance compared with baselines.

2. Task introduction

2.1. Problem definition

We regard the acronym extraction task as a sequence annotation task. Different from the previous methods, considering the high accuracy of acronym recognition, we will first recognize the acronym, and then use the character information of the acronym to recognize the long-term. Given a document $\mathbf{D} = \{x_1, x_2, \dots, x_n\}$, the initials of each word in the document is $\mathbf{I} = \{y_1, y_2, \dots, y_n\}$. Utilizing our **IRF** model, we will get each acronyms and long-term :

$$A, L = \mathbf{IRF}\left(x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_n\right) \tag{1}$$

where A refers to acronyms and L refers to long-term.

2.2. Evaluation metric

The online results will be evaluated with the macro-averaged precision, recall, and F1 scores. The final score is the prediction correctness of short-form (i.e., acronym) and long-form (i.e., phrase) boundaries in the given sentence. The short-form or long-form predictions are correct once the beginning and the end of the position of the predicted short-form or long-form are equal to the label respectively. The official score is counted based on the macro average of short-form and long-form F1 scores.

2.3. Dataset introduction

Table 1
Statistical Information of Spanish Dataset.

Data	Sample Number	Ratio
Training Set	5928	80.00%
Development Set	741	10.00%
Test Set	741	10.00%
Total	7410	100%

This task contains various multi-lingual datasets composed of document sentences in science fields. Among them, the statistics of the Spanish and the Danish datasets are shown in Table 1. The Spanish dataset is divided into training (5928), development (741), and testing (741) sets from the whole dataset. As shown in Table 2, the Danish dataset is divided into training (3082), development (385), and testing (160) sets according to the whole dataset. Both datasets have been manually labeled, where the label is a list of position boundaries.

Table 2
Statistical Information of Danish Dataset.

Data	Sample Number	Ratio
Training Set	3082	80.00%
Development Set	385	9.99%
Test Set	386	10.1%
Total	3853	100%

3. Methodology

In this section, we will introduce our proposed IRF model. IRF utilizes the corresponding relationship between the characters of acronyms and the initials of long-term, this will effectively help the model improve the accuracy of long-term recognition.

3.1. Encoder

Given a document $\mathbf{D} = \{x_1, x_2, \dots, x_n\}$, and the initials of each word in the document is $\mathbf{I} = \{y_1, y_2, \dots, y_n\}$. We leverage the pre-trained language model as an encoder to obtain the embedding as follows:

$$H = \mathbf{BERT\ Encode}\left(x_1, x_2, \dots, x_n; y_1, y_2, \dots, y_n\right) \tag{2}$$

where $H = [h_1, h_2, \dots, h_n]$ is the embedding of each token, I is the embedding of each initial.

3.2. Acronyms Tagger

The low level tagging module is designed to recognize all possible acronyms in the input sentence by directly decoding the encoded vector H produced by the N-layer BERT encoder. More precisely, it adopts two identical binary classifiers to detect the start and end position of acronyms respectively by assigning each token a binary tag (0/1) that indicates whether the current token corresponds to a start or end position of a acronym. The detailed operations of the acronyms tagger on each token are as follows:

$$p_i^{start_a} = \sigma(W_{start} h_i + b_{start}) \tag{3}$$

$$p_i^{end_a} = \sigma(W_{end} h_i + b_{end}) \tag{4}$$

where $p_i^{start_a}$ and $p_i^{end_a}$ represent the probability of identifying the i-th token in the input sequence as the start and end position of a acronym, respectively. The corresponding token will be assigned with a tag 1 if the probability exceeds a certain threshold or with a tag 0 otherwise. h_i is the encoded representation of the i-th token in the input sequence, i.e., h_i=$H[i]$, where W_{start} and W_{end} represent the trainable weight, and b_{start} and b_{end} are bias and σ is the sigmoid activation function.

3.3. Long-term Tagger

Considering that each acronym and its meaning are always connected together, we utilize Neighborhood Search Strategy to select the context near the search acronym, so as to extract the correct long-term.

The high level tagging module simultaneously identifies the long-term with respect to the acronyms obtained at lower level. As show in the Figure 2, for the acronyms *CMP*, We search for its corresponding long term in a limited context. Different from acronyms tagger directly decoding the encoded vector H, the Long-term Tagger takes the acronyms features and initial features into account as well. The detailed operations of the Long-term Tagger on each token are as follows:

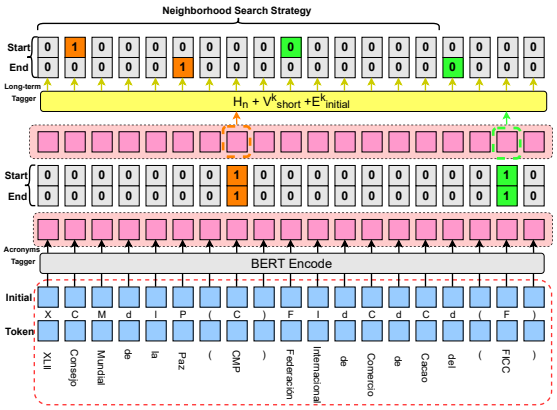

Figure 2: An overview of the proposed IRF framework. In this example, there are two candidate acronyms detected at the low level, while the presented 0/1 tags at high level are specific to the first acronym **CMP**, i.e., a snapshot of the iteration state when k = 1 is shown as above. K=2 corresponds to the second acronym **FICC**.

$$p_i^{start_l} = \sigma(W_{start}(h_n + V^k_{short} + E^k_{initial}) + b_{start}) \quad (5)$$

$$p_i^{end_l} = \sigma(W_{end}(h_n + V^k_{short} + E^k_{initial}) + b_{end}) \quad (6)$$

where $p_i^{start_l}$ and $p_i^{end_l}$ represent the probability of identifying the i-th token in the input sequence as the start and end position of a long-term respectively, and V^k_{short} represents the encoded representation vector of the k-th subject detected in low level module, the $E^k_{initial}$ represents the embedding of initials (i.e., C, M and P). For each acronym, we iteratively apply the same decoding process on it. Meanwhile, for the Neighborhood Search Strategy, we set the search length to γ, where the γ is a hyperparameter which is the longest distance between acronyms and long in the statistical training set.

4. Experiments

4.1. Baseline models

- **Rule-based method** The rule-based baseline method is proposed to adopt manual rules for this task [6]. The words with more than 60% of their characters are upper-cased to be selected as acronyms. The long-forms are chosen once the initial characters of the preceding words are before an acronym. The whole codes are online on the website[1].
- **BiLSTM-CRF model** The bidirectional LSTM [7] is an extension of LSTM that adopts a forward and backward LSTM network for sequence processing, where the links of the network is used as the output layer (Huang et al., 2015). The BiLSTM structure gathers contextual information simultaneously from the past with bidirectional. Besides, the BiLSTM has advantages in the LSTM that avoids gradient vanishing compared with the RNN. The output hidden state of BiLSTM will be concatenated between the forward LSTM H_f and backward LSTM H_b networks as final output $[H_f, H_b]$. This feature is calculated with the cross-entropy loss with the target token-level labels.
- **BERT-CRF model** The BERT-CRF [8] is implemented with the token-level neural network with the conditional random field (CRF) on top, where the backbone of this baseline can choose from the Mbert[]. The Mbert is the multilingual masked language model (MLM) trained with multiple corpora. The backbone has varients such as base and large, which are chosen as our baselines. As for the input tokens, the backbone encodes the tokens to the encoding. The final classification scores are obtained in the CRF layer, where the tag is used as the transition matrix. The matrix contains two states including the beginning (B)

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

Table 3
F1 Performance in Spanish dataset

Method	Val F1	Test F1
Rule-based	0.5667	0.5596
BiLSTM-CRF	0.7717	0.7623
BERT-CRF	0.8397	0.8211
Roberta-CRF	0.8667	0.8531
IRF-BERT$_{base}$(*ours*)	0.8742	0.8537
IRF-BERT$_{large}$(*ours*)	0.9035	0.8911
IRF-Roberta$_{large}$(*ours*)	**0.9233**	**0.91.21**

Table 4
F1 Performance in Danish dataset

Method	Val F1	Test F1
Rule-based	0.7021	0.6842
BiLSTM-CRF	0.7671	0.7587
BERT-CRF	0.8673	0.8554
Roberta-CRF	0.8979	0.8931
IRF-BERT$_{base}$(*ours*)	0.9133	0.9032
IRF-BERT$_{large}$(*ours*)	0.9532	0.9413
IRF-Roberta$_{large}$(*ours*)	**0.9744**	**0.9641**

and the end (E). This baseline is trained with the first sub-token via the cross-entropy loss.

- **Roberta-CRF model** The Roberta-CRF [9] is the same architecture as the BERT-CRF, where the difference is that the Roberta model removes the next sentence prediction (NSP) task, and uses dynamic masking for text encoding. The Roberta model uses the Byte-Pair Encoding (BPE) to mix character-level and word-level representations and support processing many common natural language corpora vocabularies. We adopt different varients of the Roberta as our baselines, including the base and the large version.

4.2. Datasets

We evaluated our method on two acronym extraction datasets mainly including Spanish dataset and Danish dataset. Specifically, the Spanish dataset has 7410 samples, and the Danish dataset has 3853 samples [10].

4.3. Implementation Detail

We used cased BERT-base, or RoBERTa-large as the encoder on Spanish and Danish dataset. All models are implemented based on the open-source transformers library of huggingface [11]. we initialize the model with mbert [12]. We use mixed-precision training [13] based on the Apex library. Our model is optimized with AdamW [14] using learning rates $\in [2e{-}5, 3e{-}5, 5e{-}5, 1e{-}4]$, with a linear warmup [15] for the first 6% steps followed by a linear decay to 0. We report the mean and standard deviation of F1 on the development set by conducting 5 runs of training using different random seeds. We utilize the In-trust loss [5] function to optimize the model.

4.4. Results

In the Spanish and Danish datasets, we compare IRF with baselines, including Rule-based, BiLSTM-CRF, BERT-CRF and Roberta-CRF. Results in Table 3 and Table 4 show that PAEE performs better than these methods. Specifically, in Spanish dataset, our best model, IRF built upon *Roberta$_l$arge*, is **+5.66 / +5.90 F1** better on Val/Test set than Roberta-CRF. In addition, in Danish dataset, IRF built upon *Roberta$_l$arge*, is **+7.65 / +7.10 F1** better on Val/Test set than Roberta-CRF. They obtain new state-of-the-art(SOTA) results, **we held the first position on the CodaLab scoreboard under the alias WENGSYX**[2].

Table 5
Test F1 score (%) on extracting long-term.

Model	Val F1	Test F1
BERT-CRF	80.42	79.11
IRF-BERT$_{base}$(*ours*)	85.31 (+ **4.89**)	84.23 (+ **5.12**)
Roberta-CRF	83.44	82.19
IRF-Roberta$_{large}$(*ours*)	90.13 (+ **6.69**)	89.07 (+ **6.88**)

4.5. Analysis

Considering the correlation between acronyms and the initials of long-term, our IRF establishes the relationship between acronyms and long-term, which improves the accuracy of extracting long and the overall performance of the model. In order to further explore the effectiveness of our method, we analyze the accuracy of identifying long-term in the acronym extraction task. As show in Table 5, compared with baseline, our IRF can significantly improve the accuracy of extracting long-term. Specifically, on the F1 score, we have a maximum performance improvement of 5%. The significant increase of the recognition accuracy of the model in long term will help to improve the overall performance of the model.

5. Conclusion

In this paper, we propose a novel Initial Reminder Framework (IRF) for acronym extraction task. Specifically,

[2]https://competitions.codalab.org/competitions/34925results

IRF utilizes Acronyms Tagger to recognize the span of acronym for the first time. Then combining with the initial information, IRF utilizes Long-term Tagger to recognize the long-term. IRF captures the relationship between acronyms and long-term in the dataset. Meanwhile, utilizing the character information in acronyms, the IRF improves the accuracy of long-term recognition. We conduct experiments on two acronym extraction datasets. Experimental results demonstrate that our IRF model can achieves state-of-the-art performance compared with baselines.

References

[1] 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).

[2] 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.

[3] 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.

[4] H. Zhao, L. Huang, R. Zhang, Q. Lu, H. Xue, Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3239–3248.

[5] 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, Association for Computational Linguistics, Online, 2021, pp. 4791–4803. URL: https://aclanthology.org/2021.findings-acl.423. doi:10.18653/v1/2021.findings-acl.423.

[6] 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.

[7] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).

[8] 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).

[9] 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).

[10] 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.

[11] 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).

[12] J. Libovický, R. Rosa, A. Fraser, How language-neutral is multilingual bert?, arXiv preprint arXiv:1911.03310 (2019).

[13] 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, in: International Conference on Learning Representations, 2018.

[14] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2018.

[15] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He, Accurate, large minibatch sgd: Training imagenet in 1 hour (2018).