Acronym Extraction with Hybrid Strategies

Siheng Li1*, Cheng Yang1*, Tian Liang1*, Xinyu Zhu1, Chengze Yu1, Yujiu Yang1†

1 Tsinghua Shenzhen International Graduate School, Tsinghua University
{lisiheng21, yangc21, liangt21, zhuxy21, ycz21}@mails.tsinghua.edu.cn,
yang.yujiu@sz.tsinghua.edu.cn

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.

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 (Ackermann et al. 2020), definition extraction (Kang et al. 2020) and relation extraction (Shi and Yang 2020; Ding et al. 2020). Recently, SDU@AAAI-22 released multiple datasets (Amir Pouran Ben Veyseh 2022a) for scientific document understanding, and we focus on the task of acronym extraction (Amir Pouran Ben Veyseh 2022b), 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 (Okazaki and Ananiadou 2006) or manual features (Kuo et al. 2009) which are labor-force and time-consumed.

* These authors contributed equally.
† Corresponding Author

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

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.

Recently, deep learning based methods (Zhu et al. 2021; Egan and Bohannon 2021) 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 (Devlin et al. 2019) or RoBERTa (Liu et al. 2019) 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.

Related Works

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

Rule-based

Traditional acronym extraction methods mainly focus on rule-based methods. Specifically, most of them (Schwartz and Hearst 2003) utilize generic rules or text patterns to discover acronym expansions in the field of biomedicine. Torres-Schumann and Schulz (2006) further extend rule sets to hidden Markov models and improve both recall and precision values. Recently, a new work (Harris and Srivinasan 2019) 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.

LSTM-based

Taking advantage of LSTM (Hochreiter and Schmidhuber 1997)’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 (Jin, Liu, and Lu 2019) extracts contextualized features with BioELMo (Jin et al. 2019) 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. (2019) 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. (2020) propose to combine LSTM with CRF for Acronym identification and Disambiguation.

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) (Devlin et al. 2018), 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 (Wei et al. 2020) and dialogue state tracking (Kim et al. 2020).

In addition, it is worth mentioning that there have been many fine-grained improvements or specific domain variants of BERT. RoBERTa (Liu et al. 2019) 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 (Beltagy, Lo, and Cohan 2019) 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. (2021) 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. (2021) utilizes Sentence Piece byte-pair encoding to relabel sentences. Then, they are embedded into the XLNet (Yang et al. 2019) for processing.

Methodology

Task Formulation

Given a text $X = \{x_1, x_2, ..., x_I\}$ where each $x_i$ is a word and $I$ 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 $A = \{[s_1, e_1], [s_2, e_2], ..., [s_n, e_n]\}$, 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 $B = \{[s_1, e_1], [s_2, e_2], ..., [s_m, e_m]\}$, similar with A.

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.

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 $X = \{x_1, x_2, ..., x_I\}$, with the help of deep multi-head attention layers, BERT captures contextualized representation for each token. The encoding process is as follows:

$$H = \text{BERT}([x_1, x_2, ..., x_I]) = [h_1, h_2, ..., h_I]^T,$$

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

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:

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

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

$$L = [L_0, L_1, L_2, L_3, L_4] = H'W_L,$$
Figure 2: The model architecture of our Sequence Labeling strategy.

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) (Sutton and McCallum 2012), the probability of a tagged sequence is:

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

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:

$$L_{SL} = -\log(P(Y|X)).$$

For the inference, we use the Viterbi algorithm (Sutton and McCallum 2012) for decoding the best label sequence.

**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 $[\text{start}, \text{end}]$ provided by raw datasets to token-level $[\text{start}, \text{end}]$ 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:

$$L = [L_0, L_1, L_2, L_3] = H'W_S,$$

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:

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

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.

**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 (Cheng et al. 2021):

$$L_{Adv} = \max_{||\epsilon||\leq \epsilon_0} \sum \text{Div}(f_0(X)||f_0(X + \epsilon)),$$
Table 1: Performance comparison on the development sets of scientific domain.

<table>
<thead>
<tr>
<th>Method</th>
<th>English</th>
<th>Persian</th>
<th>Vietnamese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>Rule</td>
<td>0.33</td>
<td>0.15</td>
<td>0.20</td>
</tr>
<tr>
<td>BERT</td>
<td>0.82</td>
<td>0.85</td>
<td>0.83</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>0.84</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>Ours-SL</td>
<td>0.86</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>Ours-SS</td>
<td>0.86</td>
<td>0.89</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 2: Performance comparison on the test sets of scientific domain, * indicates the score of our model.

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

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Training</th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>English Scientific</td>
<td>3980</td>
<td>497</td>
<td>498</td>
</tr>
<tr>
<td>Persian</td>
<td>1336</td>
<td>167</td>
<td>168</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>1274</td>
<td>159</td>
<td>160</td>
</tr>
<tr>
<td>English Legal</td>
<td>3564</td>
<td>445</td>
<td>446</td>
</tr>
<tr>
<td>French</td>
<td>7783</td>
<td>973</td>
<td>973</td>
</tr>
<tr>
<td>Spanish</td>
<td>5928</td>
<td>741</td>
<td>741</td>
</tr>
<tr>
<td>Danish</td>
<td>3082</td>
<td>385</td>
<td>386</td>
</tr>
</tbody>
</table>

Experiments

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

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.
- BERT-based This method employs BERT (Devlin et al. 2018) 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 (Liu et al. 2019) as text encoder.

Implementations
For baselines, we select pre-train models trained with corresponding language corpora in HuggingFace Transformers (Wolf et al. 2020). As for ours, we adopt the best pre-trained models according to their performance in the development set. Specifically, we adopt roberta-base for English, roberta-fa-zwnj-base-ner for Persian, bert-base-vi-

We use Jensen-Shannon divergence in our experiments.

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

$$L = L_{SL} + \alpha L_{Adv}.$$  \hfill (9)

For span selection strategy:

$$L = L_{SS} + \alpha L_{Adv}.$$  \hfill (10)

The $\alpha$ is used for controlling the significance of adversarial training.
### Table 4: Performance comparison on the development sets of legal domain.

<table>
<thead>
<tr>
<th>Method</th>
<th>English P</th>
<th>R</th>
<th>F1</th>
<th>French P</th>
<th>R</th>
<th>F1</th>
<th>Spanish P</th>
<th>R</th>
<th>F1</th>
<th>Danish P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule</td>
<td>0.32</td>
<td>0.10</td>
<td>0.16</td>
<td>0.22</td>
<td>0.06</td>
<td>0.10</td>
<td>0.17</td>
<td>0.07</td>
<td>0.10</td>
<td>0.10</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td>BERT</td>
<td>0.88</td>
<td>0.87</td>
<td>0.88</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.89</td>
<td>0.90</td>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>0.87</td>
<td>0.88</td>
<td>0.88</td>
<td>0.78</td>
<td>0.76</td>
<td>0.77</td>
<td>0.88</td>
<td>0.88</td>
<td>0.90</td>
<td>0.92</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>Ours-SL</td>
<td>0.88</td>
<td>0.88</td>
<td>0.89</td>
<td>0.95</td>
<td>0.94</td>
<td>0.94</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>Ours-SS</td>
<td>0.89</td>
<td>0.88</td>
<td>0.89</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
</tr>
</tbody>
</table>

### Table 5: Performance comparison on the test sets of legal domain.

<table>
<thead>
<tr>
<th>Ranking</th>
<th>English P</th>
<th>R</th>
<th>F1</th>
<th>French P</th>
<th>R</th>
<th>F1</th>
<th>Spanish P</th>
<th>R</th>
<th>F1</th>
<th>Danish P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.90</td>
<td>0.92</td>
<td>0.91</td>
<td>0.94</td>
<td>0.95</td>
<td>0.94</td>
<td>0.90</td>
<td>0.91</td>
<td>0.91</td>
<td>0.95</td>
<td>0.98</td>
<td>0.96</td>
</tr>
<tr>
<td>2</td>
<td>0.88*</td>
<td>0.91*</td>
<td>0.90*</td>
<td>0.92*</td>
<td>0.93*</td>
<td>0.93*</td>
<td>0.90</td>
<td>0.91</td>
<td>0.90</td>
<td>0.95</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>3</td>
<td>0.87</td>
<td>0.91</td>
<td>0.89</td>
<td>0.93</td>
<td>0.92</td>
<td>0.92</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.95*</td>
<td>0.95*</td>
<td>0.95*</td>
</tr>
<tr>
<td>4</td>
<td>0.87</td>
<td>0.90</td>
<td>0.88</td>
<td>0.81</td>
<td>0.80</td>
<td>0.81</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.89</td>
<td>0.90</td>
<td>0.89</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>English P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours-SL</td>
<td>0.86</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>Ours-SL w/o CRF</td>
<td>0.84</td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td>Ours-SL w/o AT</td>
<td>0.86</td>
<td>0.87</td>
<td>0.86</td>
</tr>
<tr>
<td>Ours-SS</td>
<td>0.86</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>Ours-SS w/o AT</td>
<td>0.86</td>
<td>0.87</td>
<td>0.86</td>
</tr>
</tbody>
</table>

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.

**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.

### 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.

### 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.

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


