Acronym Identification using Transformers and Flair Framework

F. Balouchzahi[1], O. Vitman[1], H.L. Shashirekha[2], G. Sidorov[1] and A. Gelbukh[1]

[1]Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico
[2]Department of Computer Science, Mangalore University, Mangalore, India

Abstract

The amount of acronyms in texts is growing with the increase in the number of scientific articles and it is not bound only to English texts. The Acronym Extraction (AE) task aims at automatically identifying and extracting the acronyms and their long forms in the given text. To tackle the challenge of AE in different languages, this paper describes the participation of the team MUCIC in the AE shared task at the AAAI-22 Workshop on Scientific Document Understanding (SDU@AAAI-22). This shared task aims at identifying and extracting acronyms and their long forms from English, Spanish, French, Danish, Persian, and Vietnamese texts. The proposed methodology consists of data transformation using Spacy and/or other libraries depending on the language and a Flair framework to fine-tune the transformers of the corresponding languages to extract acronyms and their long-forms. For the Spanish language, the proposed methodology secured the second rank and for all other languages, the results obtained are reasonable.

Keywords

Acronym, Expansion, Flair, BERT

1. Introduction

The term "acronym" is defined as a word or name framed by taking the first letters of each word of a phrase [1]. For instance, AIDS is an acronym for "Acquired Immune Deficiency Syndrome". Acronyms are used in a text to familiarize the abbreviations. They also serve important purposes such as speeding up the reading, avoiding repetition of unwieldy technical terms and ease the understanding of the content in a scientific paper.

Scientists frequently over-use acronyms. According to the report [2]: after an analysis of more than 24 million article titles and 18 million article abstracts published between 1950 and 2019, there was at least one acronym in 19% of titles and 73% of abstracts. This shows that the number of acronyms is constantly increasing with the increase in the amount of scientific papers published every year. Thus, the widespread usage of acronyms poses a challenge to machines and non-expert human beings attempting to read the scientific documents.

Understanding the correlation between acronyms and their expansions is critical for several Natural Language Processing (NLP) applications such as Text Classification (TC), Information Retrieval (IR) and text summarization. Therefore, it is necessary to develop a system that can automatically extract acronyms and their meanings (i.e.,long-forms or expansions) in the given documents.

Most of the existing dominant approaches to identify acronyms and their expansions in free text focus on local acronyms, whose expansions appear in the same document, typically in the same sentence or nearby sentences usually enclosed within parentheses. In contrast, non-local (global) acronyms are unaccompanied by their expansion in the same document. They are usually written with the (not necessarily correct) assumption that the reader is already familiar with the acronyms' intended meanings. Non-local acronyms are more challenging to interpret since the expansions are not found in the neighbourhood.

Over the past two decades, several techniques have been proposed to extract acronyms and their expansions from a given text corpus. These techniques use pattern-matching [3], Machine Learning (ML) (i.e., CRF and SVM) [4, 5] or word-embedding [6] to extract acronyms. More recently, Deep Learning (DL) methods [7] are showing promising results for AE. Further, pre-trained language models such as ELMo [8] and BERT [9] have also shown their effectiveness in contextual representation for extracting acronyms.

The usage of acronyms is common in many high-resource as well as low-resource languages. This paper describes the model submitted by our team MUCIC to AE shared task at SDU@AAAI-22[1] [10]. The shared task consists of identifying the acronyms and long forms from

The Second Workshop on Scientific Document Understanding at AAAI 2022

frs_b@yahoo.com (F. Balouchzahi); ovitman2021@cic.ipn.mx (O. Vitman); hlsrekha@gmail.com (H.L. Shashirekha); sidorov@cic.ipn.mx (G. Sidorov); gelbukh@cic.ipn.mx (A. Gelbukh)
https://sites.google.com/view/fazlfrs/home (F. Balouchzahi); https://www.cic.ipn.mx/~sidorov/ (G. Sidorov); http://www.gelbukh.com/ (A. Gelbukh)
0000-0003-1937-3475 (F. Balouchzahi)

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

texts in six languages, namely: English, Spanish, French, Danish, Persian, and Vietnamese.

The proposed methodology to identify acronyms in the given text contains Data transformation and Model Fine-Tuning and is based on our previous work [11] that utilized Flair framework to fine-tune transformers. Our proposed model obtained promising results for almost all high-resource languages and the best performance is achieved for Spanish with a F1-score of 0.90 leading to second rank in the AE shared task.

The rest of the paper is organized as follows: Section 2 describes some of the good performing models submitted to Acronym Identification (AI) shared task at AAAI-21 Workshop on Scientific Document Understanding (SDU@AAAI-21) followed by the proposed methodology in Section 3. Experiments and results are discussed in Section 4 and the paper concludes in Section 5.

2. Related Work

Researchers have developed several efficient models starting from traditional rule-based to advanced DL methods for AI, AE and Acronym Disambiguation (AD) tasks. Given an acronym and several possible expansions, AD task has to determine the correct expansion for the given context. AD task is challenging due to the high ambiguity of acronyms. The organizers of SDU@AAAI-21 have released two large datasets of English scientific papers published at arXiv for two shared tasks: AI [12] and AD [13]. The studies and models related to AI, AE and AD are described below:

Traditional approaches of sequence labeling, mainly rule-based or feature-based, are introduced by Schwartz et al. [14] for AI. Their model builds a dictionary of local acronyms by utilizing character-match between acronym letters and corresponding long-forms in the same sentence to discover the acronym and its long-form.

Zhu et al. [15] proposed AT-BERT - a Bidirectional Encoder Representations from Transformers (BERT)-based model for AI shared task at SDU@AAAI-21. A Fast Gradient Method (FGM)-based adversarial training strategy was incorporated in the fine-tuning of BERT variants, and an average ensemble mechanism was devised to capture the better representation from multi-BERT variants. This proposed model secured first rank in AI shared task with an average macro F1-score of 0.94.

The model proposed by Egan et al. [16] uses a transformer followed by linear projection for AI and finds similar examples with embeddings learned from Twin Networks for AD. With ensemble of different transformers, the models obtained F1-scores of 0.93 and 0.91 for AI and AD shared tasks respectively.

Pan et al. [17] introduces a binary classification model for AD. Using BERT encoder for input representations, they adopted several strategies including dynamic negative sample selection, task adaptive pretraining, adversarial training and pseudo-labeling for AD. The experiments conducted won the first place in AD shared task at SDU@AAAI-2021 with F1-score of 0.94.

Three models based on Bidirectional Long Short-Term Memory (BiLSTM) and Conditional Random Field (CRF)[2], namely: BiLSTM with CRF Huang et al. [18], Stacked BiLSTM and CRF Lample et al. [19], and Bi-LSTM and CRF with convolution and max-pooling Ma et al. [20] were adopted by Rogers et al. [21] for AI shared task with Glove embedding for all the models. They also employed four transformer models, namely: BERT, BioBERT, DistilBERT, and RoBERTa as well for AI shared task. The best performance was obtained using stacked BiLSTM with CRF with a F1-score of 0.91.

Despite several models, the complexity of AI/AE provides scope for further experimentation.

3. Methodology

The proposed methodology is adopted from our previous work on Automatic Detection of Occupations and Profession in Medical Texts using Flair and BERT [11] applied only on Spanish language texts. With minor modifications to the existing architecture, the methodology is extended for the AE task in six languages text provided by the organizers. The workflow of the methodology contains two major parts: Data Transformation and Model Fine-Tuning, which are explained in the following subsections:

3.1. Data Transformation

This phase contains the necessary steps to transform the data to a representation that can be used to train and fine-tune the model. The data provided for our previous work [11] was in Brat standoff format[3] and this data was transformed to CONLL IOB[4] format as it is easy to process data in CONLL IOB format rather than in Brat format. Brat format consists of a collection of text (.txt) and their corresponding annotation files (.ann).

The datasets for the AI shared task consists of JSON files. Each JSON file contains a collection of 4 components comprising of text, beginning and ending offsets of acronyms and their corresponding long-forms and an id of that text. A sample JSON file is shown in Figure 1. These JSON files are first transformed to Brat representation as shown in Figure 2 and then the Brat representations are transformed to CONLL IOB representation as described in [11] and is shown in Figure 3.

[2]https://github.com/guillaumegenthial/tf_ner
[3]https://brat.nlplab.org/standoff.html
[4]https://nlp.lsi.upc.edu/freeling/node/83

```
{
  "text": "WebDict 0.2919 Backoff 0.3282 Table 1: Mean Average Precision (MAP), averaged over 34 topics",
  "acronyms": [
    [
      63,
      66
    ]
  ],
  "long-forms": [
    [
      39,
      61
    ]
  ],
  "ID": "2"
},
```


Figure 1: A sample JSON file

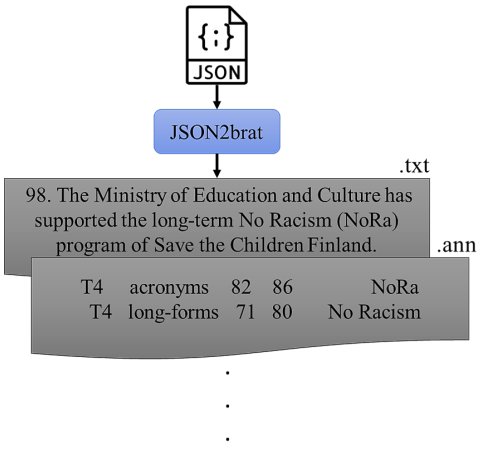

Figure 2: Transformation of data from JSON to Brat format

The input JSON files of all the languages in the given dataset are first converted to a collection of text (.txt) and their corresponding annotation files (.ann) according to Brat format based on the provided beginning and ending offsets corresponding to acronyms and their long-forms. As the proposed methodology is based on our previous work, the direct transformation of JSON files to CONLL IOB format is avoided.

Spacy[5] library which provides various tools for processing texts in different languages is used specifically to extract tokens and sentences from text. However, as Spacy does not support low resource languages such as Persian and Vietnamese, the tools pyvi[6] and HAZM[7] are used to extract tokens and sentences from Vietnamese and Persian texts respectively.

[5]https://spacy.io/
[6]https://pypi.org/project/pyvi/
[7]https://github.com/sobhe/hazm

Language	Transformer
English	bert-base-uncased
Spanish	dccuchile/bert-base-spanish-wwm-cased
Danish	Maltehb/danish-bert-botxo
French	gilf/french-camembert-postag-model
Persian	HooshvareLab/bert-fa-zwnj-base
Vietnamese	lamhieu/distilbert-base-multilingual -cased-vietnamese-topicifier

Table 1
Transformer used for each language

3.2. Model Fine-Tuning

Model Fine-Tuning employs Flair framework to fine-tune the pre-trained transformer language model to build a sequence tagger for the task of AE - a downstream task. Flair[8] is a PyTorch based NLP tool that provides the facility of utilizing individual or combination of word embeddings and language models [11]. Sequence Tagger module from Flair has BiLSTM backend with CRF layer on top of this model (which is not used in this work).

Since fine-tuning the transformers is time consuming and require significant resources such as RAM and GPU, models are fine-tuned only for 3 epochs which may probably lead to lower results. As the overall performance of the proposed methodology also depends on the language model, for each language, the most popular language model is selected and fine-tuned. The pre-trained transformer language models used for each language are presented in Table 1 and the overview of proposed methodology is shown in Figure 4.

4. Experiments and Results

The primary requirement to promote research in any NLP task is the availability of annotated dataset. AE shared task organizers have provided the participants with labeled training and development set as well as unlabelled test set for evaluating the developed models. The

[8]https://github.com/flairNLP/flair

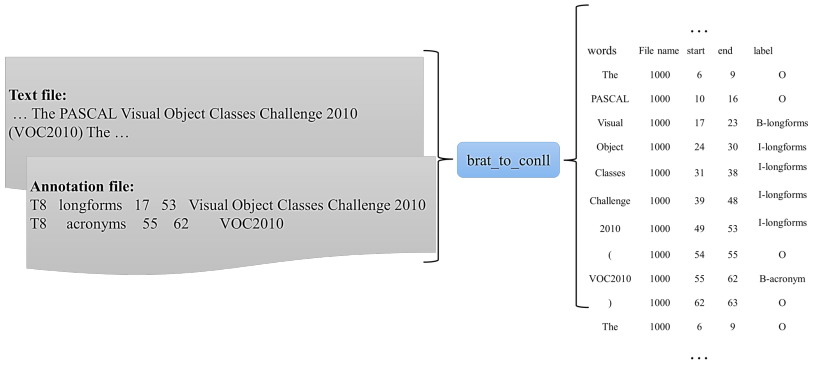

Figure 3: Transformation of data from Brat to CONLL IOB format

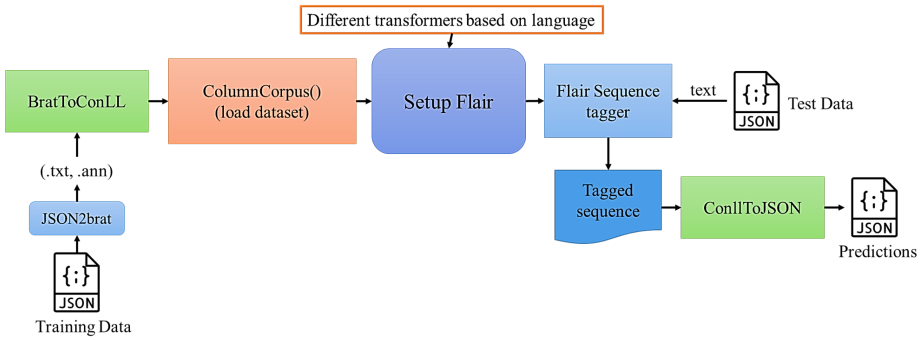

Figure 4: Overview of proposed method

datasets are provided in six languages, namely: English, Spanish, Danish, French, Persian, and Vietnamese and only English language dataset consists of legal and scientific domains [22]. Description of the datasets is available in the GitHub page[9] and their statistics are shown in Table 2. It can be observed that the datasets are highly imbalanced. Further, more number of samples in languages such as Spanish and French may lead to better performance of the task as compared to less number of samples in Vietnamese and Persian languages.

The models submitted to the shared task are evaluated on the blinded test set for predicting the boundaries of acronyms and their long-forms based on the macro-averaged scores such as Precision, Recall and F1-score. Participating teams are ranked based on macro-averaged F1-score and the results obtained by the proposed method for all languages are presented in Table 3. As expected the proposed method obtained lower results in Persian and Vietnamese languages (Spacy does not support these languages) compared to the results in other languages. The reason for lower results in Persian and Vietnamese could be due to the presence of only acronyms and their long forms in English (in some cases no long forms also) and the rest of the text in their native script. As the transformers used for these languages are monolingual, they usually do not support other scripts. The proposed model obtained its best performance in Spanish language and obtained second rank in the shared task.

Comparison of macro-averaged F1-scores of the top models in the shared task for all languages is illustrated in Figure 5. It can be observed that, as per the expectations most of models obtained higher performance in English and Spanish languages. The results also prove that as the proposed methodology with only 3 epochs training has shown promising results, experiments could be conducted on improving the results by increasing the epochs.

5. Conclusion and Future Work

This paper provides the description of the methodology and the results obtained by team MUCIC for AE shared

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

Language	Train set			Dev. set			Test set
	# of Texts	# of Acronyms	# of Long forms	# of Texts	# of Acronyms	# of Long forms	# of Texts
English (Legal)	3,563	9,532	5,246	444	1,213	669	445
English (Scientific)	3,979	7,689	5,715	469	970	720	497
Spanish	5,927	13,016	9,393	740	1,538	1,108	740
Danish	3,081	6,282	2,119	384	784	271	385
French	7,782	21,746	13,638	972	2,651	1,628	972
Persian	1,335	2,451	209	166	311	17	167
Vietnamese	1,273	1,332	62	158	175	8	159

Table 2
Statistics of the datasets used in the shared task

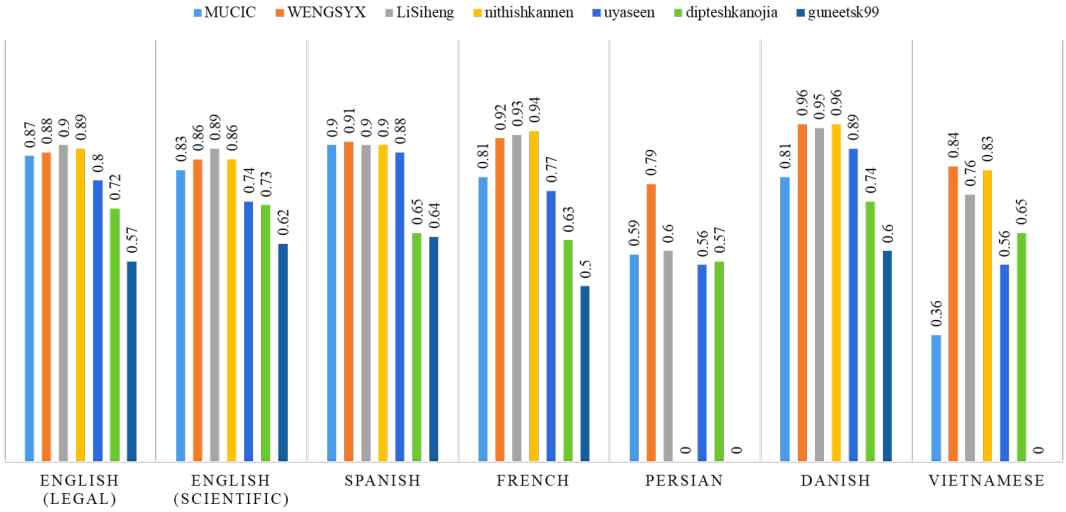

Figure 5: Comparison of macro-averaged F1-scores of top models in the shared task

task at SDU@AAAI-22. Data transformation which deals with different data representations is the primary step in this methodology. The sentences and tokens required for this step are extracted using Spacy or other libraries depending on the language. Flair framework used for fine-tuning the pre-trained transformer language model for NER task is extended by building a sequence tagger to extract acronyms and their long forms. Results obtained for different languages prove that more number of samples in the training set leads to the higher performances in identifying the acronyms and their long-forms. The proposed model obtained its best performance in Spanish language and obtained second rank in the shared task and for all other languages, the results obtained are quite reasonble. As future work we would like to experiment the combination of embeddings and language models using Flair frame work as well as other DL methods for the task of AE in different languages.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of the CONACYT, Mexico, grants 20211784, 20211884, and 20211178 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

Languages	F1-score	Precision	Recall	Rank
English (Legal)	0.87	0.84	0.89	5
English (Scientific)	0.83	0.80	0.86	5
Spanish	0.90	0.90	0.91	2
Danish	0.81	0.78	0.84	5
French	0.81	0.81	0.80	4
Persian	0.59	0.92	0.43	3
Vietnamese	0.36	0.37	0.36	6

Table 3
Performance of the proposed methodology

References

[1] C. A. Mack, How to Write a Good Scientific Paper: Acronyms, Journal of micro/nanolithography, MEMS, and MOEMS 11 (2012) 040102.

[2] A. Barnett, Z. Doubleday, The Growth of Acronyms in the Scientific Literature, eLife Sciences Publications, Ltd 9 (2020) e60080. URL: https://doi.org/10.7554/eLife.60080. doi:10.7554/eLife.60080.

[3] K. Taghva, J. Gilbreth, Finding Acronyms and their Definitions, IJDAR 1 (1999) 191–198. doi:10.1007/s100320050018.

[4] J. Liu, C. Liu, Y. Huang, Multi-Granularity Sequence Labeling Model for Acronym Expansion Identification, Information Sciences 378 (2017) 462–474.

[5] K. Jacobs, A. Itai, S. Wintner, Acronyms: Identification, Expansion and Disambiguation, Annals of Mathematics and Artificial Intelligence 88 (2020) 517–532.

[6] K. Kirchhoff, A. M. Turner, Unsupervised Resolution of Acronyms and Abbreviations in Nursing Notes using Document-level Context Models, in: Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, 2016, pp. 52–60.

[7] J. Charbonnier, C. Wartena, Using Word Embeddings for Unsupervised Acronym Disambiguation, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 2610–2619.

[8] M. Peters, M. Neumann, L. Zettlemoyer, W.-t. Yih, Dissecting Contextual Word Embeddings: Architecture and Representation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1499–1509.

[9] 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.

[10] 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.

[11] F. Balouchzahi, G. Sidorov, H. L. Shashirekha, ADOP FERT-Automatic Detection of Occupations and Profession in Medical Texts using Flair and BERT, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021, volume 2943 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 747–757. URL: http://ceur-ws.org/Vol-2943/meddoprof_paper2.pdf.

[12] A. P. B. Veyseh, F. Dernoncourt, T. H. Nguyen, W. Chang, L. A. Celi, Acronym Identification and Disambiguation Shared Tasks for Scientific Document Understanding, in: Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021, volume 2831 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2831/paper33.pdf.

[13] 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 COLING, 2020.

[14] A. Schwartz, M. Hearst, A Simple Algorithm for Identifying Abbreviation Definitions in Biomedical Text, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 4 (2003) 451–62. doi:10.1142/9789812776303_0042.

[15] 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, CEUR Workshop Proceedings (2021).

[16] N. Egan, J. Bohannon, Primer AI's Systems for Acronym Identification and Disambiguation, in: Proceedings of the Workshop on Scientific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021, volume 2831 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2831/paper30.pdf.

[17] C. Pan, B. Song, S. Wang, Z. Luo, BERT-based Acronym Disambiguation with Multiple Training Strategies, in: Proceedings of the Workshop on Sci-

entific Document Understanding co-located with 35th AAAI Conference on Artificial Inteligence, SDU@AAAI 2021, Virtual Event, February 9, 2021, volume 2831 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2831/paper25.pdf.

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

[19] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural Architectures for Named Entity Recognition, in: Proceedings of NAACL-HLT, 2016, pp. 260–270.

[20] X. Ma, E. Hovy, End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1064–1074.

[21] W. Rogers, A. R. Rae, D. Demner-Fushman, AI-NLM exploration of the Acronym Identification Shared Task at SDU@ AAAI-21., 2021.

[22] 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.