Domain Adaptive Pretraining for Multilingual Acronym Extraction

Usama Yaseen[1,2], Stefan Langer[1,2]

[1]Technology, Siemens AG Munich, Germany
[2]CIS, University of Munich (LMU) Munich, Germany

Abstract

This paper presents our findings from participating in the multilingual acronym extraction shared task SDU@AAAI-22. The task consists of acronym extraction from documents in 6 languages within scientific and legal domains. To address multilingual acronym extraction we employed BiLSTM-CRF with multilingual XLM-RoBERTa embeddings. We pretrained the XLM-RoBERTa model on the shared task corpus to further adapt XLM-RoBERTa embeddings to the shared task domain(s). Our system (team: SMR-NLP) achieved competitive performance for acronym extraction across all the languages.

Keywords

pretraining, domain adaptation, acronym extraction, XLM-RoBERTa

1. Introduction

The number of scientific papers published every year is growing at an increasing rate [1]. The authors of the scientific publications employ abbreviations as a tool to make technical terms less verbose. The abbreviations take the form of acronyms or initialisms. We refer to the abbreviated term as "acronym" and we refer to the full term as the "long form". On one hand, the acronyms enable avoiding frequently used long phrases making writing convenient for researchers but on the other hand they pose a challenge to non-expert human readers. This challenge is heightened by the fact that the acronyms are not always standard written, e.g. XGBoost is an acronym of eXtreme Gradient Boosting [2]. Following the increase of scientific publications, the number of acronyms is enormously increasing as well [3]. Thus, automatic identification of acronyms and their corresponding long forms is crucial for scientific document understanding tasks.

The existing work in acronym extraction consists of carefully crafted rule-based methods [4, 5] and feature-based methods [6, 7]. These methods typically achieve high precision as they are designed to find long form, however, they suffer from low recall [8]. Recently, Deep Learning based sequence models like LSTM-CRF [9] have been explored for the task of acronym extraction, however, these methods require large training data to achieve optimal performance. One of the major limitations of existing work in acronym extraction is that most prior work only focuses on the English language.

Woodstock'21: Symposium on the irreproducible science, June 07–11, 2021, Woodstock, NY
usama.yaseen@siemens.com (U. Yaseen);
langer.stefan@siemens.com (S. Langer)

2. Task Description and Contributions

We participate in the Acronym Extraction task [10] organized by the Scientific Document Understanding workshop 2022 (SDU@AAAI-22). The task consists of identifying acronyms (short-forms) and their meanings (long-forms) from the documents in six languages including Danish (da), English (en), French (fr), Spanish (es), Persian (fa) and Vietnamese (vi). The task corpus [11] consists of documents from the scientific (en, fa, vi) and legal domain (da, en, fr, es).

Following are our multi-fold contributions:

1. We model multilingual acronym extraction as a sequence labelling task and employed contextualized multilingual *XLM-RoBERTa* embeddings [12]. Our system consists of a single model for multilingual acronym extraction and hence is practical for real-world usage.
2. We investigated domain adaptive pretraining of *XLM-RoBERTa* on the task corpus, which resulted in improved performance across all the languages.

3. Methodology

In the following sections we discuss our proposed model for acronym extraction.

3.1. Multilingual Acronym Extraction

Our sequence labelling model follows the well-known architecture [13] with a bidirectional long short-term memory (BiLSTM) network and conditional random field (CRF) output layer [14]. In order to address the multilingual aspect of the task we employed contextualized multilingual *XLM-RoBERTa* embeddings [12] in all the experiments.

	epochs	all P/R/F1	da P/R/F1	en-sci P/R/F1	en-leg P/R/F1	fr P/R/F1	fa P/R/F1	es P/R/F1	vi P/R/F1
					dev				
r1	0	.841/.868/.854	.825/.833/.829	.727/.750/.738	.758/.784/.771	.738/.742/.740	.619/.539/.576	.820/.871/.845	.375/.547/.445
r2	1	.855/.876/.866	.826/.833/.830	.747/.757/.752	.786/.793/.789	.756/.750/.753	.644/.560/.599	.832/.872/.852	.385/.615/.474
r3	3	**.857/.878/.868**	**.827/.833/.830**	**.750/.759/.755**	**.789/.795/.792**	**.788/.751/.754**	**.665/.557/.606**	**.832/.873/.852**	**.408/.689/.512**
r4	3	-	.77/.773/.775	.617/.703/.650	.677/.677/.677	.715/.733/.724	.864/.294/.439	.823/.850/.836	.623/.074/.132
					test				
r5	3	-	.825/.833/.829	.727/.750/.738	.758/.784/.771	.738/.742/.740	.619/.539/.576	.820/.871/.845	.375/.547/.445

Table 1
F1-score on the development set (r1-r4) and test set (r5). Here, *epochs*: number of pretraining epochs for XLM-RoBERTa on the task corpus, *eng-sci*: english scientific domain, *eng-leg*: english legal domain, *all*: all languages combined.

Language	train	dev
da	3082	385
eng-scientific	3980	497
eng-legal	3564	445
fr	7783	973
es	5928	741
fa	1336	167
vi	1274	159

Table 2
Sentence counts of train and development set across the languages.

Hyperparameter	Value
hidden size	256
learning rate	$5.0e-6$
training epochs	20
pretraining epochs	3

Table 3
Hyperparameter settings for acronym extraction.

3.2. Domain Adaptive Pretraining

The original *XLM-RoBERTa* embeddings [12] are trained on the filtered CommonCrawl data (General domain), whereas the data of the shared task comprises documents from scientific and legal domains. In order to better adapt the contextualized representation to the target scientific and legal domain, we further pretrained the original XLM-RoBERTa model on the corpus data. Our experiments demonstrate improved performance on the task of acronym extraction due to the domain adaptive pretraining across all the languages.

4. Experiments and Results

4.1. Dataset

Table 2 reports sentence counts in the train and development set for all the languages. Persian and Vietnamese have substantially low sentences compared to the rest of the languages in the corpus. As a pre-processing step, we used *spaCy* [15] to perform word tokenization and POS tagging.

We do not apply any strategy to explicitly account for low training data of Persian and Vietnamese. Table 3 lists the best configuration of hyperparameters. We compute macro-averaged F1-score using the script provided by the organizers on the development set [1]. We employ early stopping and report the F1-score on the test set using the best performant model on the development set.

4.2. Results

Table 1 reports the F1-score on the development and test set for all the languages. As a baseline experiment, we combined the training data for all the languages and trained a BiLSTM-CRF model using the pretrained multilingual XLM-Roberta[2] embeddings (row r1). This achieves the overall F1-score of 0.854.

We pretrained XLM-Roberta model for 1 epoch on the task corpus using train and development set, which results in 0.1 points improvement in the overall F1-score leading to the F1-score of 0.866 (row r2). Increasing the pretraining epochs to 3 results in an improvement of additional 0.1 points in the overall F1-score (row r3).

We also experimented with training the individual models for each language (including separate models for English scientific and English legal). This results in a significant decrease in F1-score for all the languages (on average 0.12 points in F1-score, see row r4). This demonstrates that BiLSTM-CRF with multilingual XLM-Roberta embeddings performs best when trained with several languages together enabling effective cross-lingual transfer.

The F1-score of our submission on the test set are reported in row r5. Our test submission achieves the F1-score similar to the development set for all the languages demonstrating effective generalization on the test set; Vietnamese is an exception where F1-score on the test set is significantly worse than the F1-score on the development set (see rows r5 vs r3).

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

[2] https://huggingface.co/xlm-roberta-base

5. Conclusion

In this paper, we described our system with which we participate in the multilingual acronym extraction shared task organized by the Scientific Document Understanding workshop 2022 (SDU@AAAI-22). We formulate multlilignual acronym extraction in 6 languages and 2 domains as a sequence labelling task and employed BiLSTM-CRF model with multilingual XLM-RoBERTa embeddings. We pretrained XLM-RoBERTa model on the target scientific and legal domain to better adapt multilingual XLM-RoBERTa embeddings for the target task. Our system demonstrates competitive performance on the multilingual acronym extraction task for all the languages. In future, we would like to improve error analysis to further enhance our multilingual acronym extraction models.

Acknowledgments

This research was supported by the Federal Ministry for Economic Affairs and Energy (Bundesministerium für Wirtschaft und Energie: https://bmwi.de), grant 01MD19003E (PLASS: https://plass.io) at Siemens AG (Technology), Munich Germany.

References

[1] L. Bornmann, R. Mutz, Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, J. Assoc. Inf. Sci. Technol. 66 (2015) 2215–2222. URL: https://doi.org/10.1002/asi.23329.

[2] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, R. Rastogi (Eds.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, ACM, 2016, pp. 785–794. URL: https://doi.org/10.1145/2939672.2939785.

[3] A. P. A. Barnett, Z. Doubleday, The growth of acronyms in the scientific literature, eLife 9 (2020).

[4] A. S. Schwartz, M. A. Hearst, A simple algorithm for identifying abbreviation definitions in biomedical text, in: R. B. Altman, A. K. Dunker, L. Hunter, T. E. Klein (Eds.), Proceedings of the 8th Pacific Symposium on Biocomputing, PSB 2003, Lihue, Hawaii, USA, January 3-7, 2003, 2003, pp. 451–462. URL: http://psb.stanford.edu/psb-online/proceedings/psb03/schwartz.pdf.

[5] N. Okazaki, S. Ananiadou, Building an abbreviation dictionary using a term recognition approach, Bioinform. 22 (2006) 3089–3095. URL: https://doi.org/10.1093/bioinformatics/btl534.

[6] 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. URL: https://doi.org/10.1186/1471-2105-10-S15-S7.

[7] J. Liu, C. Liu, Y. Huang, Multi-granularity sequence labeling model for acronym expansion identification, Inf. Sci. 378 (2017) 462–474. URL: https://doi.org/10.1016/j.ins.2016.06.045.

[8] C. G. Harris, P. Srinivasan, My word! machine versus human computation methods for identifying and resolving acronyms, Computación y Sistemas 23 (2019). URL: https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3249.

[9] 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: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, International Committee on Computational Linguistics, 2020, pp. 3285–3301. URL: https://doi.org/10.18653/v1/2020.coling-main.292.

[10] A. P. B. Veyseh, N. Meister, S. Yoon, R. Jain, F. Dernoncourt, T. H. Nguyen, Multilingual Acronym Extraction and Disambiguation Shared Tasks at SDU 2022, in: Proceedings of SDU@AAAI-22, 2022.

[11] A. P. B. Veyseh, N. Meister, S. Yoon, R. Jain, F. Dernoncourt, T. H. Nguyen, MACRONYM: A Large-Scale Dataset for Multilingual and Multi-Domain Acronym Extraction, in: arXiv, 2022.

[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747.

[13] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: K. Knight, A. Nenkova, O. Rambow (Eds.), NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, The Association for Computational Linguistics, 2016, pp. 260–270. URL: https://doi.org/10.18653/v1/n16-1030.

[14] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Condi-

tional random fields: Probabilistic models for segmenting and labeling sequence data, in: C. E. Brodley, A. P. Danyluk (Eds.), Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, Morgan Kaufmann, 2001, pp. 282–289.

[15] M. Honnibal, I. Montani, M. Honnibal, H. Peters, S. V. Landeghem, M. Samsonov, J. Geovedi, J. Regan, G. Orosz, S. L. Kristiansen, P. O. McCann, D. Altinok, Roman, G. Howard, S. Bozek, E. Bot, M. Amery, W. Phatthiyaphaibun, L. U. Vogelsang, B. Böing, P. K. Tippa, jeannefukumaru, GregDubbin, V. Mazaev, R. Balakrishnan, J. D. Møllerhøj, wbwseeker, M. Burton, thomasO, A. Patel, explosion/spaCy: v2.1.7: Improved evaluation, better language factories and bug fixes, 2019. URL: https://doi.org/10.5281/zenodo.3358113. doi:10.5281/zenodo.3358113.