MACRONYM: A Large-Scale Dataset for Multilingual and Multi-Domain Acronym Extraction

Amir Pouran Ben Veyseh[1], **Nicole Meister**[2], **Seunghyun Yoon**[3],
Rajiv Jain[3], **Franck Dernoncourt**[3], and **Thien Huu Nguyen**[1]

[1]Department of Computer Science, University of Oregon, OR, USA
[2]Department of Computer Science, Stanford University, CA, USA
[3]Adobe Research, USA

`{apouranb, thien}@cs.uoregon.edu, nmeist@stanford.edu`
`{syoon,rajijain,franck.dernoncourt}@adobe.com`

Abstract

Acronym extraction is the task of identifying acronyms and their expanded forms in texts that is necessary for various NLP applications. Despite major progress for this task in recent years, one limitation of existing AE research is that they are limited to the English language and certain domains (i.e., scientific and biomedical). Challenges of AE in other languages and domains are mainly unexplored. As such, lacking annotated datasets in multiple languages and domains has been a major issue to prevent research in this direction. To address this limitation, we propose a new dataset for multilingual and multi-domain AE. Specifically, 27,200 sentences in 6 different languages and 2 new domains, i.e., legal and scientific, are manually annotated for AE. Our experiments on the dataset show that AE in different languages and learning settings has unique challenges, emphasizing the necessity of further research on multilingual and multi-domain AE.

1 Introduction

Acronyms are short forms of longer phrases that are often constructed using a few letters selected from the long phrases. Due to their functionality, acronyms are common in many languages and domains. For instance, 73% of abstracts of scientific papers contain at least an acronym (Barnett and Doubleday, 2020). As such, in text processing applications, e.g., question answering and machine translation, it is necessary to correctly identify the acronyms and their meanings. Toward this goal, our work focus on the task of Acronym Extraction (AE), aiming to recognize acronyms and their definitions/long forms in text. For instance, in the sentence "*They will meet in the conference of the World Trade Organization (WTO)*", an AE system should identify "*WTO*" and "*World Trade Organization*" as the acronym and long form, respectively.

Despite all progress in recent years, prior work on AE has mainly limited to specific domains and languages. Specifically, biomedical and scientific texts in English have been the main focus in prior work. However, recognition of acronyms in other languages and domains is also important and might involve challenges not reflected in English biomedical/scientific texts. For instance, many existing AE methods for English employ uppercase letters to identify acronyms (Veyseh et al., 2020). However, in non-case sensitive languages, e.g., Arabic or Persian, uppercase letter concept does not exist, thus causing a failure of existing AE systems. Moreover, in each domain or language, different styles might be exerted to shorten a longer phrase to produce acronyms. For instance, initial letters of the words in the phrases are commonly used to form acronyms in scientific English; however, in legal English or Danish documents, the use of initial letters for acronym detection is less effective (see Section 3). As such, it is desirable to study AE in more diverse domains and languages to better support multi-domain and multilingual applications.

Unfortunately, to the best of our knowledge, there is no existing dataset for multilingual and multi-domain AE, thus impeding research effort in this area. To this end, our work addresses this issue by introducing a new manually labeled dataset for AE. In particular, based on two different domains of scientific and legal texts, our dataset annotates AE data for sentences in six different languages: English, Danish, Spanish, French, Persian, and Vietnamese. As such, legal texts, Danish, Spanish, French, Persian, and Vietnamese are not explored for AE in prior work. In addition, our dataset is large-scale, providing 27,200 annotated sentences for AE to support advanced model development (e.g., with data-hungry deep learning models).

Finally, we conduct extensive experiments to understand the challenges of AE in the created dataset. Our experiments show that the AE task in our dataset presents significant challenges for existing models in different domains and languages.

This is even more pronounced in the cross-lingual and cross-domain transfer settings where existing models perform poorly on our AE dataset. As such, more research effort is needed to address the challenges of acronym understanding in different settings. We will publicly release the dataset to foster research in this area.

2 Data Annotation

Data Collection: We collect data in two domains of legal and scientific documents for AE annotation. For each domain, documents in different languages are required. As such, for the legal domain, we employ the United Nations Parallel Corpus (UNPC) (Ziemski et al., 2016) and the Europarl corpus (Koehn, 2005). The UNPC corpus contains official records in 6 languages while the Europarl corpus consists of the proceedings of the European Parliament in European languages. To accommodate our annotation budget and diversify the resulting dataset, we choose documents from four languages in the two corpora (i.e., English, French, and Spanish in UNPC, and Danish in Europarl) for our AE annotation. In addition, for the scientific domain, we employ the publicly available papers and M.S./Ph.D. theses in the field of computer science for AE annotation. Specifically, we collect the papers published in the ACL anthology of natural language processing research for English. Also, for typologically different languages, we crawl public computer science thesis in Persian and Vietnamese.

Following (Veyseh et al., 2020), we split the selected documents into sentences that will be annotated separately by annotators. In addition, to optimize the annotation cost with greater numbers of acronyms, we apply the same procedure in (Veyseh et al., 2020) to filter out sentences that has low chance to contain acronyms or long forms. In particular, the procedure only retains sentences that involve at least one acronym candidate (i.e., a word with more than a half of characters as capital letters) and a sub-sequence of words to match the acronym candidate (i.e., concatenating the initials of the words can form the candidate) (Veyseh et al., 2020). Here, we only apply this procedure for English, French, Spanish, and Danish as our Persian and Vietnamese data is small and the sentence filtering procedure will leave less sentences for annotation. Finally, given the retained sentences for each language, we randomly sample a subset of sentences for manual AE annotation. The numbers

of annotated sentences are presented in Table 1.

Annotation Process: To annotate the sampled sentences, we recruit native speakers in each language from the crowd-sourcing platform `upwork.com` with freelancer annotators across the globe. For each language, we select annotator candidates who have experience in related annotation projects and an approval rate of more than 95% (provided by Upwork). The annotator candidates are trained with guidelines and examples for AE in their language. In our annotation guideline, acronyms are required to be single words (including abbreviations). Also, for a sentence in a language, we only annotate the long forms that are in the same language as the sentence's. Afterward, for each language, we retain two candidates who pass and achieve highest results in our designed test for AE as our official annotators. Next, the two annotators in each language independently perform AE annotation for the sampled sentences of that language. Finally, the two annotators will discuss to resolve any disagreement in the annotation, thus producing a final version of our MACRONYM dataset.

	Domain & Language	IAA	Size	# Unique Acronyms	# Unique Long-forms
Legal	English	0.824	4,000	3,688	3,037
	Spanish	0.810	6,400	4,059	4,437
	French	0.823	8,000	5,638	5,728
	Danish	0.810	3,000	907	923
Scientific	English	0.811	4,000	3,604	4,260
	Persian	0.782	1,000	641	203
	Vietnamese	0.791	800	270	61

Table 1: Statistics of MACRONYM. IAA scores use Krippendorff's alpha with MASI distance based on initial independent annotations. Size refers to the number of annotated sentences.

To study the challenges of AE in each language, following (Veyseh et al., 2020), we compute the inter-annotator agreement (IAA) scores using Krippendorff's alpha (Krippendorff, 2011) with the MASI distance metric (Passonneau, 2006) for the initial independent annotations of the two annotators, i.e., before resolving the conflicts. Table 1 shows the IAA scores for each language. Overall, we find that the IAA scores are high for all considered languages and domains, thus demonstrating the quality of our annotated dataset. Among several factors, a major scenario of annotation disagreement occurs in Persian or Vietnamese when a sentence contains a long form term that is translated from an original English term. However, the acronym for this long form in the Persian or Viet-

namese sentence is still formed via initials of the words in the English term. As such, some annotators consider this English-based acronym as an acronym in the Persian or Vietnamese sentence while other annotators simply ignore it in the annotation. For instance, in the Persian sentence:

"عملیات پیشرفته شبکه (ANOP) به شرح زیر است"[1],

the acronym "*ANOP*" is expressed in English letters but its long form, i.e., "عملیات پیشرفته شبکه"[2], is presented in Persian. In the resolving, we have decided to annotate any acronym that is formed using characters in the six languages in our dataset.

Data Analysis: We show the main statistics of MACRONYM in Table 1. This table shows that the density of acronyms in texts varies across different languages. On average, English sentences tend to involve more acronyms than other languages in both legal and scientific domain while Danish and Vietnamese sentences contain least acronyms in the legal and scientific domain respectively. Comparing English texts in the legal and scientific domains, we find that the ratio between the numbers of unique long-forms and acronyms is greater in the scientific domain, thus implying the higher ambiguity of acronyms in scientific documents. Finally, we note that the number of unique acronyms exceeds the number of unique long forms in Persian and Vietnamese as we do not apply the sentence filtering procedure in the data collection, thus allowing many sentences with only acronyms and no associated long forms to be annotated in the data.

3 Experiments

This section studies the challenges of the multilingual and multi-domain AE task in MACRONYM. In particular, for each pair of available languages and domains (we have 7 pairs in total), we first prepare the data by randomly splitting the corresponding set of annotated sentences into separate training/development/test portions with the ratios of 80/10/10 (respectively). Afterward, we report the performance of the representative AE models on the test set for each possible pair of languages and domains under different learning settings.

AE Models: We examine the performance of three representative state-of-the-art (SOTA) models for AE. First, we employ the rule-based system for AE proposed in (Veyseh et al., 2021) (called

Rule-Based). This system serves as the current SOTA rule-based method for AE (Veyseh et al., 2021). In general, to detect acronyms, words with more than 60% characters as uppercase letters are selected. To find long-forms, if a detected acronym is bounded between parentheses and the initial letters of preceding words can form the acronym, the system predicts the preceding words a long form. Second, motivated by prior work (Veyseh et al., 2020; Zhu et al., 2021), we solve AE as a sequence labeling problem using BIO tagging schema. In particular, following the current SOTA deep learning model for AE (Zhu et al., 2021), we employ a pretrained BERT-based language model followed by a feed-forward network layer with softmax in the end to predict BIO-based label for each word in the sentence. To facilitate the learning on multiple languages, we explore two multilingual transformer-based language models, i.e., mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2020), leading to two models **mBERT** and **XLMR** for this approach.

Settings: MACRONYM enables the evaluation of AE models on four different settings: (i) **Mono-Lingual Mono-Domain**: In this setting, training and test data of the models come from the same language and domain. As we have 7 possible pairs of languages and domains, this setting involves 7 different evaluations for each AE model; (ii) **Mono-Lingual Cross-Domain**: Training and test data for models belongs to the same languages, but different domains in this setting. In MACRONYM, this setting is only possible for English where AE models are trained on the legal domain but tested on the scientific domain and vice versa (i.e., two possible evaluations).; (iii) **Cross-Lingual Mono-Domain**: Assuming the same domain for training and test data, this setting trains models on English training data and evaluate them on test data of other languages. We thus have 3 and 2 possible evaluations for the legal and scientific domains respectively.; (iv) **Cross-Lingual Cross-Domain**: Training and test data for models originates from different languages and domains in this setting. As such, we also consider five evaluations in this setting. In the first two evaluations, models are trained on English data in the legal domain and evaluated on Persian and Vietnamese test data in the scientific domains. In contrast, for the other three evaluations, English data in the scientific domain is used for model training while Spanish, French, and Danish test data in the legal domain is used for evaluation.

---

[1]English translation: "*Advanced network operations (ANOP) include the followings*"

[2]English translation: "*advanced network operations*"

Domain & Language		Mono-Lingual Mono-Domain			Mono-Lingual Cross-Domain		Cross-Lingual Mono-Domain		Cross-Lingual Cross-Domain	
		Rule-Based	mBERT	XLMR	mBERT	XLMR	mBERT	XLMR	mBERT	XLMR
Legal	English	16.55	61.66	62.07	54.92	56.88	-	-	-	-
	Spanish	10.82	51.43	55.41	-	-	38.88	40.13	35.48	36.92
	French	10.05	58.77	61.14	-	-	48.82	50.70	44.21	46.83
	Danish	8.78	50.05	48.38	-	-	40.71	42.94	38.18	41.95
Scientific	English	20.72	60.51	59.00	56.71	59.88	-	-	-	-
	Persian	60.59	62.41	63.10	-	-	49.13	50.21	42.95	43.72
	Vietnamese	53.44	58.71	59.13	-	-	50.72	51.44	48.32	50.17

Table 2: Model performance (F1 scores) in different settings. The performance in each row is evaluated on the test data for the corresponding pair of language and domain. **Mono-Lingual Cross-Domain**: trained on English data of one domain and tested on English data of the other domain; **Cross-Lingual Settings**: trained on English data of one domain and tested on the other language data of the same domain (if Mono-Domain) or the other domain (if Cross-Domain).

We fine-tune the hyper-parameters for the models using the development data for each pair of languages and domains. Our fine-tuning process returns similar values of hyper-parameters for the models across languages and domains. In particular, for English legal data, we use 2 layers of feed-forward neural networks with 200 hidden dimensions for hidden vectors. The learning rate is set to $3e$-4 for the Adam optimizer and the batch size of 8 is utilized for training.

Results: Table 2 presents the performance of three AE models in four different settings. Note that as the **Rule-Based** system does not require training, its performance in the mono-lingual and mono-domain setting can be applied to other settings. There are several observations from the table. First, the **Rule-Based** system achieves decent performance for Persian and Vietnamese, but performs poorly for other languages. The main reason has to do with the dominance of acronyms over long forms in Persian and Vietnamese data (see Table 1). This is in contrast to other languages where acronyms and long forms are more balanced. As acronyms can be identified more easily with rules than long forms, the **Rule-Based** system is more effective in the data with much more acronyms of Persian and Vietnamese. Second, in the legal domain where long forms are better presented, the performance of the models on English is significantly better than those for other languages, thus demonstrating the more challenging nature of non-English language for AE. Third, compared to deep learning models, the significant lower performance of the **Rule-Based** model in the monolingual and monodomain setting signifies the brittleness of human-designed rules for AE that necessitates learning

models to improve the portability of models to different languages and domains. Fourth, across all learning models and language-domain pairs for test data, the lower performance in the cross-lingual mono-domain setting compared to its mono-lingual counterpart suggests the difference between languages that hinder cross-lingual transfer learning for AE. Fifth, the cross-domain performance also under-performs their mono-domain counterpart for almost all learning models and language-domain pairs for testing, thus highlighting domain shifts as an important challenge for AE. Finally, across all the learning settings, the performance of the AE models is still far from being perfect in MACRONYM, thus presenting ample opportunities for future research.

4 Related Work

Early attempts for AE have employed rule-based methods (Park and Byrd, 2001; Wren and Garner, 2002; Schwartz and Hearst, 2002; Adar, 2004; Nadeau and Turney, 2005; Kirchhoff and Turner, 2016) or feature engineering models (Kuo et al., 2009; Liu et al., 2017; Li et al., 2018). Recently, deep learning methods have delivered SOTA performance for AE (Veyseh et al., 2021; Wu et al., 2015; Antunes and Matos, 2017; Charbonnier and Wartena, 2018; Ciosici et al., 2019; Jaber and Martínez, 2021; Li et al., 2021). Despite such progress, prior AE research and datasets have mainly focused on English biomedical and scientific texts, leaving non-English languages and other domains less explored. Here, we note that there exist some acronym glossaries for non-English languages (Pomares-Quimbaya et al., 2020; Ménard and Ratté, 2011). However, such resources do not

annotate sentences/texts in multiple languages and domains for AE as we do in MACRONYM.

5 Conclusion

We present the first multilingual and multi-domain dataset for AE, involving annotation for 6 languages and 2 domains. Our experiments show that the proposed dataset presents significant challenges for AE methods in different learning settings and languages. In the future, we will expand the dataset to include more domains and languages for AE.

References

Eytan Adar. 2004. Sarad: A simple and robust abbreviation dictionary. *Bioinformatics*, 20(4):527–533.

Rui Antunes and Sérgio Matos. 2017. Biomedical word sense disambiguation with word embeddings. In *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pages 273–279. Springer.

Adrian Barnett and Zoe Doubleday. 2020. Meta-research: The growth of acronyms in the scientific literature. In *Elife*.

Jean Charbonnier and Christian Wartena. 2018. Using word embeddings for unsupervised acronym disambiguation. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel R Ciosici, Tobias Sommer, and Ira Assent. 2019. Unsupervised abbreviation disambiguation. *arXiv preprint arXiv:1904.00929*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Areej Jaber and Paloma Martínez. 2021. Participation of uc3m in sdu@ aaai-21: A hybrid approach to disambiguate scientific acronyms. In *SDU@ AAAI*.

Katrin Kirchhoff and Anne M Turner. 2016. 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*, pages 52–60.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Cheng-Ju Kuo, Maurice HT Ling, Kuan-Ting Lin, and Chun-Nan Hsu. 2009. Bioadi: a machine learning approach to identifying abbreviations and definitions in biological literature. In *BMC bioinformatics*, volume 10, page S7. Springer.

Feng Li, Zhensheng Mai, Wuhe Zou, Wenjie Ou, Xiaolei Qin, Yue Lin, and Weidong Zhang. 2021. Systems at sdu-2021 task 1: Transformers for sentence level sequence label. In *SDU@ AAAI*.

Yang Li, Bo Zhao, Ariel Fuxman, and Fangbo Tao. 2018. Guess me if you can: Acronym disambiguation for enterprises. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1308–1317, Melbourne, Australia. Association for Computational Linguistics.

Jie Liu, Caihua Liu, and Yalou Huang. 2017. Multi-granularity sequence labeling model for acronym expansion identification. *Information Sciences*, 378:462–474.

Pierre André Ménard and Sylvie Ratté. 2011. Classifier-based acronym extraction for business documents. In *Knowledge and information systems*.

David Nadeau and Peter D Turney. 2005. A supervised learning approach to acronym identification. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 319–329. Springer.

Youngja Park and Roy J Byrd. 2001. Hybrid text mining for finding abbreviations and their definitions. In *Proceedings of the 2001 conference on empirical methods in natural language processing*.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation. In *LREC*.

Alexandra Pomares-Quimbaya, Pilar López-Úbeda, Michel Oleynik, and Stefan Schulz. 2020. Leveraging pubmed to create a specialty-based sense inventory for spanish acronym resolution. In *Digital Personalized Health and Medicine*.

Ariel S Schwartz and Marti A Hearst. 2002. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Biocomputing 2003*, pages 451–462. World Scientific.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Walter Chang, and Thien Huu Nguyen. 2021. MadDog: A web-based system for acronym identification and disambiguation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Quan Hung Tran, and Thien Huu Nguyen. 2020. What does this acronym mean? introducing a new dataset for acronym identification and disambiguation. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Jonathan D Wren and Harold R Garner. 2002. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of information in medicine*, 41(05):426–434.

Yonghui Wu, Jun Xu, Yaoyun Zhang, and Hua Xu. 2015. Clinical abbreviation disambiguation using neural word embeddings. In *Proceedings of BioNLP 15*, pages 171–176.

Danqing Zhu, Wangli Lin, Yang Zhang, Qiwei Zhong, Guanxiong Zeng, Weilin Wu, and Jiayu Tang. 2021. At-bert: Adversarial training bert for acronym identification winning solution for sdu@aaai-21. In *Proceedings of the AAAI-21 Workshop on Scientific Document Understanding, Shared Task (SDU@AAAI-21)*.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*.