BLAR: Biomedical Local Acronym Resolver

William Hogan[1,2], **Yoshiki Vazquez Baeza**[2], **Yannis Katsis**[3], **Tyler Baldwin**[3],
Ho-Cheol Kim[3], **Chun-Nan Hsu**[2]
[1]Department of Computer Science & Engineering,
[2]Center for Microbiome Innovation
University of California, San Diego, La Jolla, CA 92093
[3]IBM Research-Almaden, 650 Harry Road, San Jose, CA 95120
whogan@ucsd.edu

Abstract

NLP has emerged as an essential tool to extract knowledge from the exponentially increasing volumes of biomedical texts. Many NLP tasks, such as named entity recognition and named entity normalization, are especially challenging in the biomedical domain partly because of the prolific use of acronyms. Long names for diseases, bacteria, and chemicals are often replaced by acronyms. We propose Biomedical Local Acronym Resolver (BLAR), a high-performing acronym resolver that leverages state-of-the-art (SOTA) pre-trained language models to accurately resolve local acronyms in biomedical texts. We test BLAR on the Ab3P corpus and achieve state-of-the-art results compared to the current best-performing local acronym resolution algorithms and models.

1 Introduction

In the past decade, natural language processing (NLP) has greatly advanced in the biomedical domain. Given the troves of biomedical texts, NLP has emerged as a critical tool for knowledge extraction. NLP has been used to automatically analyze clinical notes, electronic medical records, biological literature, and other biomedical texts in the hopes of unearthing new knowledge and deeper insights.

Acronyms are especially common in science and even more so in biomedical publications, as authors regularly seek to shorten the long names for diseases, bacteria, and chemicals. Barnett and Doubleday (2020) documented acronym use in more than 24 million scientific article titles and 18 million scientific articles published between 1950 and 2019. They report that 19% of titles and 73% of abstracts contain acronyms. Of the more than one million unique acronyms in their data, 0.2% appeared regularly and most acronyms, 79%, appeared less than 10 times.

Acronym resolution (AR) can be performed by either leveraging acronym definitions found in the text (referred to as *local AR*) or by consulting external resources, such as ontologies (known as *disambiguation* or *global AR*). While a lot of progress has been recently done on the latter, local AR has seen surprisingly little recent work. In particular, the SOTA approaches in local AR are rule-based or simple machine learning approaches from more than a decade ago. As a result, this task has not benefited from recent advances in transformers (Vaswani et al., 2017). To address this issue, in this work we focus on local AR where we try to answer the question: Can transformers be leveraged to further improve traditional local AR approaches?

To answer this question, we present Biomedical Local Acronym Resolver (BLAR); a transformer-based model designed to resolve local acronyms in biomedical texts. In particular, this work makes the following contributions:

1. *Design of a novel transformer-based model for local acronym resolution*, which resolves acronyms through a combination of a two-step architecture and appropriate leveraging of pre-trained language models. To the best of our knowledge, this is the first transformer-based approach for local AR.

2. *Experimental evaluation of BLAR against SOTA local AR approaches*, showing that it outperforms the latter. In particular, evaluated on the Ab3P corpus (Sohn et al., 2008), BLAR reaches an F1 score of 0.966 compared to 0.899 of the best performing existing approach.

2 Background and Related Work

There are a few challenges inherent in acronym resolution that make a simple dictionary-lookup and

Proceedings of the BioNLP 2021 workshop, pages 126–130
June 11, 2021. ©2021 Association for Computational Linguistics

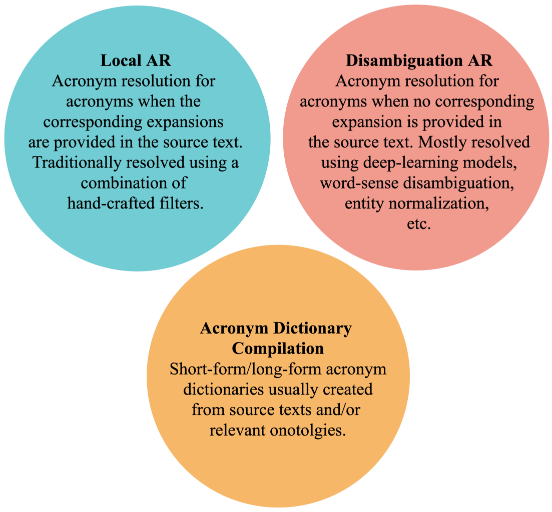
Local AR
Acronym resolution for acronyms when the corresponding expansions are provided in the source text. Traditionally resolved using a combination of hand-crafted filters.
Disambiguation AR
Acronym resolution for acronyms when no corresponding expansion is provided in the source text. Mostly resolved using deep-learning models, word-sense disambiguation, entity normalization, etc.
Acronym Dictionary Compilation
Short-form/long-form acronym dictionaries usually created from source texts and/or relevant onotolgies.

Figure 1: Sub-tasks of acronym resolution (AR). Our approach is applicable to both "Local AR" and "Acronym Dictionary Compilation."

other rule-based models less effective. First, short-form acronym representations are rarely unique. For instance, "CD" is an acronym for "Crohn's disease" and "Cowden Disease." A simple dictionary lookup of "CD" using an acronym disease dictionary will produce ambiguous results and requires additional steps of acronym disambiguation. Moreover, the number of letters in a short-form may not match the number of words in the corresponding long-form (e.g. the short-form of "systemic sclerosis" is "SSc"). Lastly, long-form entities can have complicated short-forms. For example, the short-form of "heparin-induced thrombocytopenia type II" is "HIT type II," a short-form that shortens the first three words of the long-form and leaves the last two words unmodified.

To address these challenges, approaches to acronym resolution have been developed and can be classified into three broad categories: *local* acronym resolution (Schwartz and Hearst, 2003; Sohn et al., 2008), *disambiguation* acronym resolution (also referred to as *non-local* or *global* acronym resolution) (Jin et al., 2019; Jacobs et al., 2020), and *acronym dictionary compilation* (Grossman et al., 2018). We refer to approaches that resolve acronyms by leveraging their definitions found in the containing text as *local* acronym resolution techniques. In contrast, *non-local* or *global* techniques resolve acronyms by using external resources. These typically target acronyms whose long-form is not contained within the text, which is common among more established acronyms, such as "mRNA" and "DNA." Finally, *acronym dictionary compilation* refers to the creation of an acronym dictionary based on the source text or external ontologies, or a combination of the two. These three sub-categories of AR approaches are depicted in Figure 1.

Our approach specifically targets local acronym resolution and acronym dictionary compilation. Local acronyms appear as a pair of entities featuring a short-form (SF) entity and a corresponding long-form (LF) entity. Historically, local acronym resolution has been handled by rule-based algorithms. From 2003 to 2009, Schwartz et al. (2003) and Sohn et al. (2008) demonstrated the best performance of local acronym resolution. They used a combination of hand-crafted filters to identify SF-LF pairs. Kuo et al. (2009) introduced the first local acronym resolution model that leveraged machine learning. It produced SOTA results with the help of four sets of hand-crafted features, including rule-based text filters. Yeganova et al. (2011) further improved upon local acronym resolution by introducing a hybrid machine learning and rule-base model that does not rely on labeled data. They extract potential SF-LF pairs from PubMed articles using rules similar to the rules developed by Sohn et al. and train a classifier to identify SF-LF pairs.

Our approach to local acronym resolution is simple in its architecture yet novel in its application. Our two-stage model leverages transfer learning from modern, SOTA pretrained transformers and is able to learn the features of short-form and long-form acronym pairs without the help of a predefined dictionary, hand-crafted features, filters, or rules. Our model processes batches of documents, such as abstracts from PubMed, and creates an acronym dictionary specific to each inputted document.

3 Method

The intuition behind local acronym resolution is that authors of scientific publications commonly define the acronyms that they employ later on in the document. This is typically done by defining acronyms within the text in the form of pairs of short-form (SF) and corresponding long-form (LF) entities. We can then use the identified SF-LF acronym pairs to either resolve the acronyms appearing in the input document or populate an SF-LF dictionary that can be used to accurately resolve future uses of the SF versions of the acronyms in the remainder of the text.

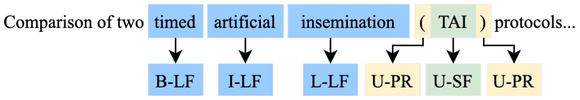

Figure 2: Sample output of *Step 2* showing the various tagged entities of a short and long-form acronym pair. We use a BILOU (Beginning, Inside, Last, Outside, Unit) tagging scheme (Ratinov and Roth, 2009) to identify long-form (LF) entities, short-form (SF) entities, and parenthesis (PR) enclosing a paired SF or LF entity.

Identifying the definitions of SF-LF pairs poses two major challenges: First, one has to identify the location in the text where the definition of an SF-LF pair is provided. Second, one has to identify the exact span (i.e., text) of both the short and long-form within the definition.

Two-step AR: Following the above structure, BLAR splits the problem into two separate subtasks:

- *Step 1: Sentence Classification.* Given the input text, identify sentences containing definitions of SF-LF pairs. This is modeled as a binary classification task.
- *Step 2: SF-LF Acronym Tagging:* Given a sentence predicted to contain a definition of an SF-LF pair, identify the exact form (i.e., text) of the SF and LF entities. This is modeled as a token classification task, where each token in the sentence is classified as being part of an acronym short-form, acronym long-form, or the parenthesis enclosing a paired entity. Token classification follows the BILOU (Beginning, Inside, Last, Outside, and Unit) encoding scheme (Ratinov and Roth, 2009), as shown in Figure 2 through a simple example.

Model architecture: The sentence classification model (*Step 1*) leverages transfer learning by fine-tuning the pretrained SciBERT model (Beltagy et al., 2019) for the specific task of sentence classification. The sentences that have been predicted as containing SF-LF pairs are given as input to the SF and LF tagging model (*Step 2*). The tagging model also leverages SciBERT by fine-tuning it on the SF and LF tagging task. To avoid exposure bias resulting from training on a set of perfect inputs (e.g. sentences containing acronym pairs as labeled in the dataset), we use the output from the sentence classification model from *Step 1* to train the tagging model in *Step 2*. The output of the tagging model is a dictionary that can then be used to replace all the short-form acronyms with their corresponding long-forms within a single source text.

Model training: We developed BLAR using the BioADI corpus (Kuo et al., 2009) and tested it on the Ab3P corpus (Sohn et al., 2008). BioADI includes 1,668 true SF-LF pairs from 1,200 annotated PubMed abstracts and Ab3P includes 1,221 true SF-LF pairs from 1,250 annotated PubMed abstracts. Both provide span-level data identifying short and long-form acronym pairs within PubMed abstracts and differ only in the articles selected for annotation. During development, we fine-tuned both our sentence and acronym token classifiers on the BioADI corpus randomly split into three subsets for training (80% of the corpus), validation (10% of the corpus), and testing (10% of the corpus). We use BioADI as a training dataset and Ab3P as a testing dataset to best compare our model's performance to existing SOTA benchmarks for local acronym resolution which use the same train/test splits. The BioADI and Ab3P corpora are described in Section 4. Since the models in both steps are fine-tuned versions of SciBERT, they are able to train fairly quick on CPUs. *Step 1* and *Step 2* converged within eight epochs, taking roughly 10 hours and 2 hours to complete, respectively, on two Intel Xeon CPUs (E5-2640 v3 @ 2.60GH) with 16GB of RAM.

Ablation study: To determine the importance of the 2-step architecture, we conduct an ablation study where we train a model to resolve acronyms without the help of a sentence classification step. This model is identical to the tagging model used in *Step 2*, only, it is trained on raw sentences that may or may not contain an acronym pair. This single-step architecture must simultaneously learn to detect and resolve an acronym pair. We refer to this model variation as “BLAR (single step).”

4 Datasets

BioADI: We use the BioADI (Kuo et al., 2009) corpus to train BLAR. It includes 1,668 true SF-LF pairs from 1,200 annotated PubMed abstracts.

Ab3P: We use the Ab3P (Sohn et al., 2008) corpus for testing. It includes 1,221 true SF-LF pairs from 1,250 annotated PubMed abstracts.

At the time of writing, both datasets are available for download on the BioC (Comeau et al., 2013) website.

5 Results and Discussion

To measure BLAR's performance, we first compare it against SOTA local AR approaches. As explained in the *Background and Previous Work* section, to the best of our knowledge, local acronym resolution has not seen significant advances since 2009. More recent acronym resolution works have focused instead on disambiguation acronym resolution, still relying on simpler rule-based algorithms for local acronym resolution (Jin et al., 2019; Jacobs et al., 2020). As a result, we compare BLAR to Kuo et al. (2009), Sohn et al. (2008), and Schwartz and Hearst (2003), which represent the SOTA in local acronym resolution.

Table 1 depicts the performance of BLAR against SOTA AR models. In this experiment, all models were trained on the BioADI dataset and tested on the Ab3P dataset. For each model, we evaluate Precision, Recall, and F1 score based on exact matches of long-form and short-form pairs. The results show that BLAR significantly outperforms all previous approaches, achieving an F1 score of 0.966 compared to 0.899 of the next best approach. We observe that, without a sentence classification step, the single-step BLAR model under-performs compared to the two-step architecture, highlighting the benefit of the sentence classification step in the full two-step architecture.

AR Model	P	R	F1
Schwartz et al. (2003)	0.950	0.788	0.861
Sohn et al. (2008)	**0.970**	0.836	0.898
Kuo et al. (2009)	0.959	0.846	0.899
Yeganova et al. (2011)	0.936	0.893	0.914
BLAR (single step)	0.950	0.957	0.953
BLAR (two step)	0.966	**0.966**	**0.966**

Table 1: Evaluation results of BLAR against SOTA local acronym resolution models. All models, save Yeganova et al., were trained on BioADI and tested on Ab3P. Yeganova et al. is trained on 1M automatically extracted potential SF-LF pairs from PubMed abstracts.

Model Output Analysis: Finally, to further understand the performance of BLAR, we perform an instance-level analysis of its output.

Analyzing the correct predictions, we see that the model successfully overcomes some of the complex challenges inherent in acronym resolution. For example, it correctly resolves the acronyms "SSc" to "systemic sclerosis" and "IUAG" to "intrauterine growth retardation." These examples show that BLAR learns to resolve short-forms that contain a different number of letters compared to the number of words in the corresponding long-form. In another example, BLAR correctly resolves "HIT type II" to "heparin-induced thrombocytopenia type II" which illustrates that the model was able to learn more complex acronyms that consist of a mix of short-form entities and complete words.

Moving to the incorrect predictions, we classify BLAR's errors into three categories: missed acronyms (false negatives), added acronyms (false positives), and modified acronyms (i.e., acronyms where the model correctly identifies a short-form but either truncates or extends the corresponding long-form).

A majority of the errors come from modified acronyms. Analyzing the modified acronyms, we find that 63.7% of cases are long-forms expanded or truncated by a single word/token. We identify that many of the erroneously expanded long-forms add a word or words preceding the ground truth long-form. For example, in the text "... heat stroke by reducing iNOS-dependent nitric oxide (NO)... ", BLAR identified "iNOS-dependent nitric oxide" as the long-form expansion of the short-form "NO.", instead of the correct "nitric oxide."

Another common error within the modified acronyms category is a truncated long-form. For example, BLAR predicts the long-form of "FVC" to be "forced vital capacity" but the ground truth is "forced expiratory volume in 1 s vital capacity." Here, BLAR predicts a simple long-form when the ground truth long-form is actually more complex. We plan to explore these insights in future work to further improve the model.

6 Conclusion and Future Work

Local acronym resolution has seen limited progress in recent years and has not benefited from the recent advancements in machine learning approaches. To address this problem, we develop BLAR; a deep-learning model that leverages a two-step architecture on top of pre-trained language models to identify SF-LF pairs in input documents. Our experimental results show that BLAR outperforms other local acronym resolution approaches and achieves state-of-the-art performance. We release BLAR and its source code for public use. As part of our future work, we will be exploring two threads: first, further improving the model based on our error analysis, and second, exploring how BLAR (which in this case has been fine-tuned for the scientific domain) can be extended to cover acronyms found in other domains. We believe future work could also focus on a hybrid model that leverages both deep-learning and rule-based algorithms.

Acknowledgement

Thank you to the anonymous reviewers for their thoughtful comments and corrections. This work is supported by IBM Research AI through the AI Horizons Network.

References

Adrian Barnett and Zoe Doubleday. 2020. Meta-research: The growth of acronyms in the scientific literature. *eLife*, 9:e60080.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Donald C. Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, Alfonso Valencia, Karin Verspoor, Thomas C. Wiegers, Cathy H. Wu, and W. John Wilbur. 2013. Bioc: a minimalist approach to interoperability for biomedical text processing. *Database : the journal of biological databases and curation*, 2013:bat064–bat064. 24048470[pmid].

Lisa V. Grossman, Elliot G. Mitchell, George Hripcsak, Chunhua Weng, and David K. Vawdrey. 2018. A method for harmonization of clinical abbreviation and acronym sense inventories. *Journal of Biomedical Informatics*, 88:62 – 69.

Kayla Jacobs, Alon Itai, and Shuly Wintner. 2020. Acronyms: identification, expansion and disambiguation. *Annals of Mathematics and Artificial Intelligence*, 88(5):517–532.

Qiao Jin, Jinling Liu, and Xinghua Lu. 2019. Deep contextualized biomedical abbreviation expansion. *Proceedings of the 18th BioNLP Workshop and Shared Task*.

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. *BMC Bioinformatics*, 10(15):S7.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 451–62.

Sunghwan Sohn, Donald C. Comeau, Won Kim, and W. John Wilbur. 2008. Abbreviation definition identification based on automatic precision estimates. *BMC bioinformatics*, 9:402–402. PMC2576267[pmcid].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lana Yeganova, Donald Comeau, and W. Wilbur. 2011. Machine learning with naturally labeled data for identifying abbreviation definitions. *BMC bioinformatics*, 12 Suppl 3:S6.