



**UNIVERSITY
OF TURKU**

This is a self-archived – parallel-published version of an original article. This version may differ from the original in pagination and typographic details. When using please cite the original.

LICENSE	Creative Commons Attribution 4.0 International License.
TITLE	m-Networks: Adapting the Triplet Networks for Acronym Disambiguation
YEAR	2022
VERSION	Publisher PDF
CITATION	Sandaru Seneviratne, Elena Daskalaki, Artem Lenskiy, and Hanna Suominen. 2022. m-Networks: Adapting the Triplet Networks for Acronym Disambiguation-Networks: Adapting the Triplet Networks for Acronym Disambiguation . In <i>Proceedings of the 4th Clinical Natural Language Processing Workshop</i> , pages 21–29, Seattle, WA. Association for Computational Linguistics.

m-Networks: Adapting the Triplet Networks for Acronym Disambiguation

Sandaru Seneviratne¹, Elena Daskalaki¹, Artem Lenskiy¹, Hanna Suominen^{1,2}

¹The Australian National University (ANU) / Canberra, ACT, Australia

²University of Turku / Turku, Finland

{sandaru.seneviratne, eleni.daskalaki,
artem.lenskiy, hanna.suominen}@anu.edu.au

Abstract

Acronym disambiguation (AD) is the process of identifying the correct expansion of the acronyms in text. AD is crucial in natural language understanding of scientific and medical documents due to the high prevalence of technical acronyms and the possible expansions. Given that natural language is often ambiguous with more than one meaning for words, identifying the correct expansion for acronyms requires learning of effective representations for words, phrases, acronyms, and abbreviations based on their context. In this paper, we proposed an approach to leverage the triplet networks and triplet loss which learns better representations of text through distance comparisons of embeddings. We tested both the triplet network-based method and the modified triplet network-based method with *m* networks on the AD dataset from the SDU@AAAI-21 AD task, CASI dataset, and MeDAL dataset. F scores of 87.31%, 70.67%, and 75.75% were achieved by the *m* network-based approach for SDU, CASI, and MeDAL datasets respectively indicating that triplet network-based methods have comparable performance but with only 12% of the number of parameters in the baseline method. This effective implementation is available at https://github.com/sandaruSen/m_networks under the MIT license.

1 Introduction

Natural language is often ambiguous and contains phrases, words, acronyms, and abbreviations which have more than one meaning (Charbonnier and Wartena, 2018). The complexity of natural language is further augmented based on which context these words are being used (Navigli, 2009). Scientific and medical communities use domain specific technical terms, which are often shorthanded for ease of use. This has resulted in the prevalence of acronyms in scientific and medical documents (Charbonnier and Wartena, 2018). To understand these expert texts, it is important to disambiguate

the meaning of their acronyms. For example, given a sentence with the acronym *RNN*, the possible expansion for the acronym can be *Recurrent Neural Network*, *Random Neural Network*, *Recursive Neural Network*, *Reverse Nearest Neighbour*, etc. Out of these expansions, the one corresponding to the meaning of the sentence should be identified in order to correctly understand the sentence. The task of identifying the correct expansion of acronyms from possible expansions is called *Acronym Disambiguation* (AD).

Methods of pattern matching, language modeling, and machine/deep learning have shown promising results in AD. Early systems for AD used pattern matching (Schwartz and Hearst, 2002) together with approaches based on word embeddings and machine learning (Jaber and Martínez, 2021) where the AD task is considered as a classification problem. Recent efforts in AD mainly include the use of deep learning-based models (Pan et al., 2021; Zhong et al., 2021) and pre-trained language models (Beltagy et al., 2019; Devlin et al., 2019). However, identifying the correct expansion of an acronym calls for better representation of text.

In this study, we approached the problem of AD with the aim of learning effective text representations towards better disambiguation of acronyms. We derived our approach from *Siamese Networks* (Koch et al., 2015) and *Triplet Networks* (TNs) (Hoffer and Ailon, 2015). TNs, inspired by Siamese Networks, aim to learn the information of inputs based on one or a few samples of training data using a triplet loss to provide better representations for data.

The main contributions of this paper were as follows: We leveraged the triplet loss and TNs (Schroff et al., 2015) for AD with the aim of learning sentence embeddings, which can capture the semantic differences of the different expansions of the same acronym. We extended the TN architecture further to include *m* networks and mapped the

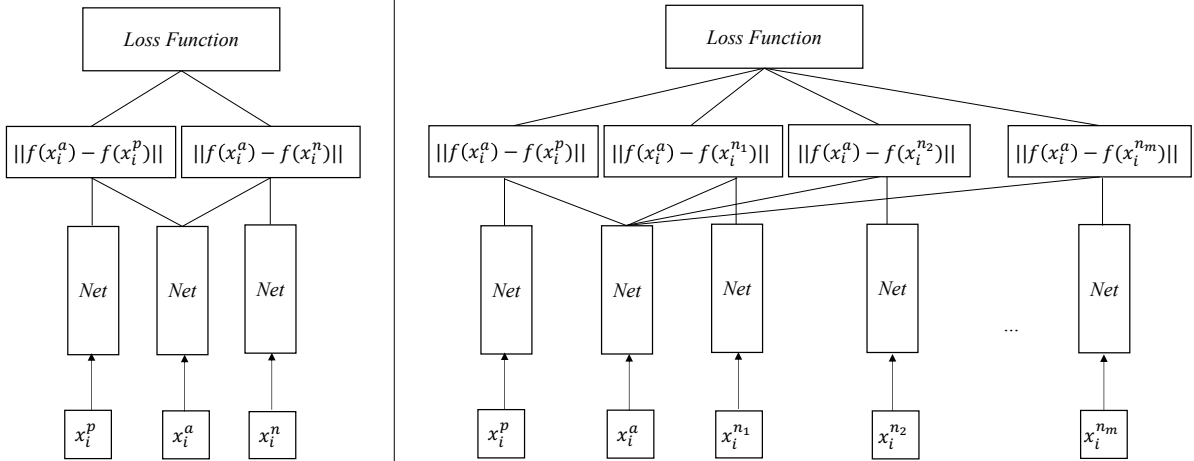


Figure 1: **Triplet Network Architecture and Modified Triplet Network Architecture.** The triplet network architecture (left, Formula (1)) considers the anchor sentence x_i^a , positive sentence x_i^p , and negative sentence x_i^n for a sample when computing the triplet loss. Modified architecture (right, Formula (2)) considers the anchor sentence, positive sentence, and all the possible negative sentences for a sample. This includes m number of similar architectures.

AD task as a binary classification problem, which predicts if the suggested expansion for an acronym is correct or not. To the best of our knowledge this is the first attempt of adapting the TN-based methods and triplet loss for disambiguating the acronyms. We evaluated and verified the proposed approach on the AAAI-21 Scientific Document Understanding AD task dataset (SDU dataset) (Veyseh et al., 2020), sense inventory for clinical abbreviations and acronym dataset (CASI dataset) (Moon et al., 2014), and on a sample of the Medical Abbreviation Disambiguation Dataset (MeDAL) (Wen et al., 2020). We made our implementation available at https://github.com/sandaruSen/m_networks under the MIT license.

2 Related Work

Extensive body of prior research for AD in scientific and medical domains exists because understanding scientific and medical text requires both AD and domain knowledge. Earliest approaches for AD included the use of a number of rules and patterns (Schwartz and Hearst, 2002), training of classifiers based on a set of features which represent the context of the input like, part-of-speech tags, case representation of the words, or word stems (Finley et al., 2016; Wu et al., 2017), and computation of the cosine similarity between the text with the acronym and the possible expansions based on word embeddings (Tulkens et al., 2016). Recent efforts in AD include the use of deep learning-based methods and pre-trained language

models (Pan et al., 2021; Singh and Kumar, 2021; Zhong et al., 2021).

With the introduction of transformers, the transformer-based pre-trained language models have been extensively used for the AD task. BERT (Bidirectional Encoder Representations from Transformers) models such as (Devlin et al., 2019), SciBERT (BERT-based language model for performing scientific tasks) (Beltagy et al., 2019), and RoBERTa (Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019) are the language models that are exploited to formulate the problem of AD as a classification task for AD. The SDU@AAAI-21 AD task consisted of systems with transformer-based language models, which differed based on how the inputs and the outputs to the systems were defined (Veyseh et al., 2021). In our work, we explored triplet loss and TNs for AD using pre-trained language models. TNs and triplet loss have been effectively used for representation learning by distance comparisons among pairs of examples. They were initially introduced for computer vision related tasks (Schroff et al., 2015) and are now used in many natural language processing (NLP) tasks (Santos et al., 2016; Ein-Dor et al., 2018; Lauriola and Moschitti, 2020; Wei et al., 2021). We believe that through the triplet loss, the models will be able to learn subtle yet complex differences among the different expansions of the same acronym.

3 Methods

The goal of AD was to identify the correct expansion for a given acronym in text. Considering a dictionary of acronyms D with acronyms as keys $[A_1, A_2, \dots, A_j]$ where j is the number of acronyms. For each acronym A_i , the m possible expansions were represented as $[e_1, e_2, \dots, e_m]$. Given a sentence x_i with an acronym A_i , the correct expansion should be obtained from D out of the expansion list of the corresponding A_i .

We modeled the AD task based on a TN as well as a modified version of the TN architecture with the triplet loss. The TN allowed the AD task to be expressed as a binary classification problem to predict which expansion is the most relevant to the given acronym based on the context it appears (Appendix A). For the modified version of the TN, we included m number of architectures considering the possible negatives for a sample at once. This resulted in an anchor sentence, a positive sentence, and a list of negative sentences as inputs to the architectures (Figure 1).

Denoting anchor, positive, and negative embeddings as x_i^a , x_i^p , and x_i^n , respectively, where $i = 1, 2, \dots, k$, and considering a d -dimensional embedding in the vector space $f(x) \in \mathbb{R}^d$ and α a margin that is enforced between positive and negative pairs, the loss for the TN was defined as follows using the L_2 distances for the TN:

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2. \quad (1)$$

For the modified version of the TN with m networks, the loss was computed considering all the possible negatives. Adapting the triplet loss to the modified architecture, the distance between the anchor and the positive sentence should be less than the minimum of the distances between the anchor and the negative sentences. We could denote the loss considering all the m number of negatives x_i^{n1} , x_i^{n2} , \dots , x_i^{nm} as follows:

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \min(\|f(x_i^a) - f(x_i^{n1})\|_2^2, \|f(x_i^a) - f(x_i^{n2})\|_2^2, \dots, \|f(x_i^a) - f(x_i^{nm})\|_2^2). \quad (2)$$

Sentence triplet creation, which includes identifying an anchor sample x_i^a , a positive sample x_i^p , and a negative sample x_i^n (Table 1), was considered crucial when using TNs. For each possible expansion of an acronym, we randomly extracted one

sentence matching the expansion from the training dataset. These sentences were considered as anchor sentences. We then used all sentences in the training dataset to create positive samples. Acronyms in sentences were replaced by their respective correct expansion to obtain positive sentences. We then applied the following guidelines to create the negative samples: i) For each positive sentence with an acronym, we obtained all the possible expansions except for the correct expansion. ii) We replaced the acronym in the sentence with these expansions to obtain a list of sentences with other expansions. iii) Each of these negative sentences was used to create the final list of triplets.

The triplet selection ensured effective training of the models. Hence, it is advised to consider triplets, which violate the triplet constraint (Formula (1)). In our approach, we considered the same positive sentence with the respective acronym replaced by other expansions of the acronym as negatives. Even though the text in the sentences was very much similar to each other, replacing the acronym with possible expansions resulted in a change in the semantic meaning of the overall sentences. Hence, we believe considering sentences with other possible expansions as negative sentences satisfied the necessity of having *hard negatives*, which were difficult to discriminate from the correct expansion.

Anchor Sentence	The purpose of <i>RL</i> is for the agent to learn an optimal, or nearly-optimal, policy that maximizes the reward function.
Positive Sentence	All agents can then operate in parallel, allowing one to exploit a number of already available <i>reinforcement learning</i> techniques for parallel learning.
Negative Sentences	[All agents can then operate in parallel, allowing one to exploit a number of already available <i>robust locomotion</i> techniques for parallel learning., All agents can then operate in parallel, allowing one to exploit a number of already available <i>representation learning</i> techniques for parallel learning., ...]

Table 1: An example of anchor, positive, and negative sentences for the acronym *RL* and the expansion *reinforcement learning*.

Architecture or Model	Number of Parameters	F score on SDU	F score on CASI	F score on MeDAL
Baseline method by Singh and Kumar (2021)	109,920,002	84.24%	78.16%	74.91%
Triplet Network-based method	13,576,768	85.70%	56.49%	75.19%
<i>m</i> Network-based method	13,576,768	87.31%	70.67%	75.75%

Table 2: Results of the validation data of SDU dataset and test data of CASI and MeDAL datasets.

In the training stage, we used the anchor sentence, positive sentence, and negative sentence as the input to the TN-based system and anchor sentence, positive sentence, and possible negative sentences as the input to the *m*-network-based system. For each of the sentences, we obtained an embedding, which was then used to calculate the triplet loss. In the inference stage, we used the given sentence with the acronym as the anchor sentence and we created a list of sentences by replacing the acronym in the sample sentence with possible expansions. We computed the distances between each of the possible sentences and the anchor sentence to obtain the sentence closest to the anchor sentence.

4 Experiments

We used the SDU dataset (Veyseh et al., 2020), CASI dataset (Moon et al., 2014), and MeDAL dataset (Wen et al., 2020) (see Appendix B for further information). The SDU dataset contained data from 6,786 English scientific papers published at arXiv and consisted of 62,441 sentences. The dataset also consisted of a dictionary of acronyms and their possible expansions. We used the publicly available training and development data of the SDU dataset for our experiments. CASI dataset was created using admission notes, consultation notes, and discharge summaries from hospitals affiliated with the University of Minnesota. 37,500 samples from CASI dataset was split into train, validation, and test subsets and a dictionary with the acronyms was created for the experiments. The MeDAL dataset was created from 14,393,619 articles in PubMed. We created a sample dataset and a dictionary of acronyms from MeDAL dataset for experiments (Table 3 of Appendix B).

We performed a basic preprocessing on the sentences, which were quite long, by sampling tokens in the sentences as proposed by Singh and Kumar (2021). We used $N/2$ tokens to the left and right of the acronym for sentences with length of more than 120, considering $N = 120$.

As a *baseline model*, we experimented with the system proposed by Singh and Kumar, 2021 which modeled the AD task as a span prediction task. The proposed system fine-tuned the complete SciBERT model with 12 layers to predict the start and end indices of the correct expansion of an acronym given all the possible expansions, leveraging the SciBERT’s ability to encode pair of sequences together.

We used the pre-trained SciBERT model architecture as the base model for experiments on SDU dataset and the pre-trained BioBERT (BERT-based language model for performing biomedical tasks) (Lee et al., 2020) model as the base model for experiments on the CASI and the MeDAL datasets with their first 11 encoder layers frozen followed by dropout of 0.5 to avoid over-fitting and a dense layer to map the feature embeddings output by the base models with dimensions of 768 to 64 (Appendix C). These 64 dimensional embeddings were used to compute the triplet loss. We trained the models using a learning rate of 5×10^{-4} with the *Adam optimizer* (Kingma and Ba, 2014). The best model over 10 epochs with a batch size of 32 was chosen as the final model.

To evaluate the performance of the proposed architecture in the training set, we computed the macro-averaged F1 score. If the distance between the anchor and the positive sentence is less than the distance between the anchor and negative sentences, the prediction of the model was considered correct. We used F1 also in evaluation. We computed the distances between the anchor and possible sentences from which the sentence with the minimum distance to the anchor was considered the sentence with the correct expansion.

5 Results and Analysis

By comparing the proposed methods with the baseline system on the three datasets, we observed that the methods based on TNs learnt to discriminate among the different expansions of an acronym. Compared to the TN-based method, the *m* network-

based method has comparable performance as the baseline for all the datasets. Both the proposed methods outperformed the baseline on SDU and MeDAL datasets. The m network-based method gave an F1 score of 87.31% on SDU dataset, 70.67% on CASI dataset, and 75.75% on MeDAL dataset (Table 2).

To investigate the semantic similarity and the representation of the output embeddings in the vector space, we visualized output representations obtained by the m network-based architecture for the SDU, CASI, and MeDAL datasets by reducing the dimensions using *principal component analysis* (PCA) (Figure 3 of Appendix D). For the SDU dataset, we used the acronym *RL* with *reinforcement learning* to obtain the positive and respective negative sentences. Similarly, for the CASI dataset the acronym *DM* with *diabetes mellitus* expansion and for the MeDAL dataset the acronym *RSM* with *respiratory muscle strength* expansion were used.

6 Discussion

In this paper, we have suggested a new approach for disambiguating the acronyms to effectively identify the correct expansion through better representation learning using TNs by creating high quality sentence embeddings, which can capture the semantic differences among the different expansions of the same acronym. Namely, we have presented how methods based on TNs and triplet loss can be used for AD. To address the effective learning of context representations for identifying the correct expansion of acronyms, our methods leverage the contextual information of text and semantic similarity among expansions. In particular, our paper has introduced m networks inspired by TNs. Our experiments have demonstrated that methods based on TNs have comparable performance on both scientific and medical domains. However, the applicability of the proposed methods on CASI dataset should be further investigated. Finally, the number of parameters in TN-based methods is only 12% of the number of parameters in the baseline method resulting in smaller size of the models (Table 2). The TN-based methods have used the representations from the last layer of the BERT-based models where as the baseline method fine-tuned the complete model with all 12 layers for the predictions¹.

¹However, given that m network-based method consists of m architectures, the number of updates on parameters increases.

We have tested the proposed methods on the SDU, CASI, and MeDAL datasets.

The TN-based method for AD can be used for data augmentation when the training data is limited. Given that the original TN architecture only considers one negative sample at a time, considering all the possible expansions of each acronym one at a time can be used to augment the training data size. This addresses the issue of limited training data for deep learning architectures. However, in the modified TN-based architecture with m networks, at the training stage all the possible negatives are considered for a sample at once. Therefore, data augmentation is not possible in this case.

In this paper, our main goal was to approach the AD problem as an effective representation learning problem to discriminate among the possible expansions of an acronym based on the context it appears. Earliest approaches on AD relied on rules and patterns (Schwartz and Hearst, 2002) to identify the correct expansion of an acronym which evolved to use of machine learning-based approaches with different features (Finley et al., 2016; Wu et al., 2017) and computing of semantic similarity between the text with acronym and the possible expansions. Recent efforts involved pre-trained language models for the AD task. Most of these systems were validated on one domain of focus (i.e., scientific text, medical text, or general text). We approached the problem focusing on learning better representations for text through TNs and triplet loss using pre-trained language models. Furthermore, we tested the proposed approaches on both the scientific and medical domains.

As future work, we intend to experiment with different contrastive losses (Sohn, 2016; Chen et al., 2020). Specifically, our aspiration is to compare and contrast the proposed approach with InfoNCE (Van den Oord et al., 2018), a popular contrastive loss which includes multiple negatives and normalises across examples in a mini batch.

7 Ethical Considerations

We have proposed an approach for AD using TN-based methods with the aim of learning effective representations for data. We have used SciBERT trained on scientific publications and BioBERT trained on biomedical domain corpora (PubMed abstracts and PMC full-text articles) for our experiments. Instead of finetuning all the layers in the pre-trained language models, we have finetuned

only the last encoder layer by freezing the first 11 encoder layers thereby bringing the latest deep learning advances to AD in a computationally efficient way. However, the m network architecture despite its smaller number of parameters has m architectures. This has resulted in more updates in the parameters increasing the computational time in the training stage.

The proposed approaches have been tested and validated on three datasets: SDU dataset, CASI dataset, and MeDAL dataset. According to the National Statement on Ethical Conduct in Human Research (2007) — Updated 2018 (National Health and Medical Research Council, 2018), a new ethics approval is not required for our experiments and, to the best of our knowledge, the three original datasets have been created ethically. All the three datasets are publicly available (see Appendix B).

Identifying the correct expansion of acronyms is important in improving the understandability of scientific/medical text due to the prevalence of technical acronyms which are shorthanded for ease of use. For people with limited expertise knowledge, understanding scientific/medical documents can be difficult, stressful and cause misunderstandings. The proposed methods can be used in scientific/medical text simplification tasks to provide lay people with better understanding of text through the disambiguation of acronyms.

Acknowledgement

This research was funded by and has been delivered in partnership with Our Health in Our Hands (OHIOH), a strategic initiative of the ANU, which aims to transform health care by developing new personalized health technologies and solutions in collaboration with patients, clinicians and health-care providers. We gratefully acknowledge the funding from the ANU School of Computing for the first author's PhD studies.

References

- 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.
- Jean Charbonnier and Christian Wartena. 2018. Using word embeddings for unsupervised acronym disambiguation. In *Proceedings of the 27th International Conference on Computational Linguistics*.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.
- 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.
- Liat Ein-Dor, Yosi Mass, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov, and Noam Slonim. 2018. Learning thematic similarity metric using triplet networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), Melbourne, Australia*, pages 15–20.
- Gregory P Finley, Serguei VS Pakhomov, Reed McEwan, and Genevieve B Melton. 2016. Towards comprehensive clinical abbreviation disambiguation using machine-labeled training data. In *AMIA Annual Symposium Proceedings*, volume 2016, page 560. American Medical Informatics Association.
- Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer.
- Areej Jaber and Paloma Martínez. 2021. Participation of uc3m in sdu@ aaai-21: A hybrid approach to disambiguate scientific acronyms. In *SDU@ AAAI*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.
- Ivano Lauriola and Alessandro Moschitti. 2020. Context-based transformer models for answer sentence selection. *arXiv preprint arXiv:2006.01285*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

- Sungrim Moon, Serguei Pakhomov, Nathan Liu, James O Ryan, and Genevieve B Melton. 2014. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, 21(2):299–307.
- National Health and Medical Research Council. 2018. National Statement on Ethical Conduct in Human Research (2007). <https://www.nhmrc.gov.au/about-us/publications/national-statement-ethical-conduct-human-research-2007-updated-2018>. [Online; accessed 06-January-2022].
- Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.
- Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. Bert-based acronym disambiguation with multiple training strategies. *arXiv preprint arXiv:2103.00488*.
- Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.
- 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.
- Aadarsh Singh and Priyanshu Kumar. 2021. Scidr at sdu-2020: Ideas-identifying and disambiguating everyday acronyms for scientific domain. In *In SDU@AAAI-21*.
- Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.
- Stéphan Tulkens, Simon Šuster, and Walter Daelemans. 2016. Using distributed representations to disambiguate biomedical and clinical concepts. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*.
- Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807.
- Amir Pouran Ben Veyseh, Franck Dernoncourt, Thien Huu Nguyen, Walter Chang, and Leo Anthony Celi. 2021. Acronym identification and disambiguation shared tasks for scientific document understanding. In *In SDU@AAAI-21*.
- 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*, pages 3285–3301.
- Jason Wei, Chengyu Huang, Soroush Vosoughi, Yu Cheng, and Shiqi Xu. 2021. Few-shot text classification with triplet networks, data augmentation, and curriculum learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5493–5500.
- Zhi Wen, Xing Han Lu, and Siva Reddy. 2020. Medal: Medical abbreviation disambiguation dataset for natural language understanding pretraining. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 130–135.
- Yonghui Wu, Joshua C Denny, S Trent Rosenbloom, Randolph A Miller, Dario A Giuse, Lulu Wang, Carmelo Blanquicett, Ergin Soysal, Jun Xu, and Hua Xu. 2017. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (card). *Journal of the American Medical Informatics Association*, 24(e1):e79–e86.
- Qiwei Zhong, Guanxiong Zeng, Danqing Zhu, Yang Zhang, Wangli Lin, Ben Chen, and Jiayu Tang. 2021. Leveraging domain agnostic and specific knowledge for acronym disambiguation. In *SDU@AAAI*.

A Triplet Networks and Triplet Loss

Triplet loss uses anchor, positive, and negative samples to learn effective representations. Anchor sample comes from a specific class. Positive samples belong to the same class as the anchor sample and the negative samples belong to a different class than the class of the anchor sample. The triplet loss encourages to minimize the distance between similar embeddings (i.e., anchor and positive embeddings) and maximize the distances between dissimilar embeddings (anchor and negative embeddings) enforcing a margin between the embeddings.

B Data Samples and Their Availability

The datasets used in this study are all publicly available from the following sources: **AD dataset from SDU@AAAI21**, **CASI**, and **MeDAL**. The dataset statistics are shown in Table 3. The distribution of the number of samples based on the number of acronym expansion pairs is shown in Figure 2.

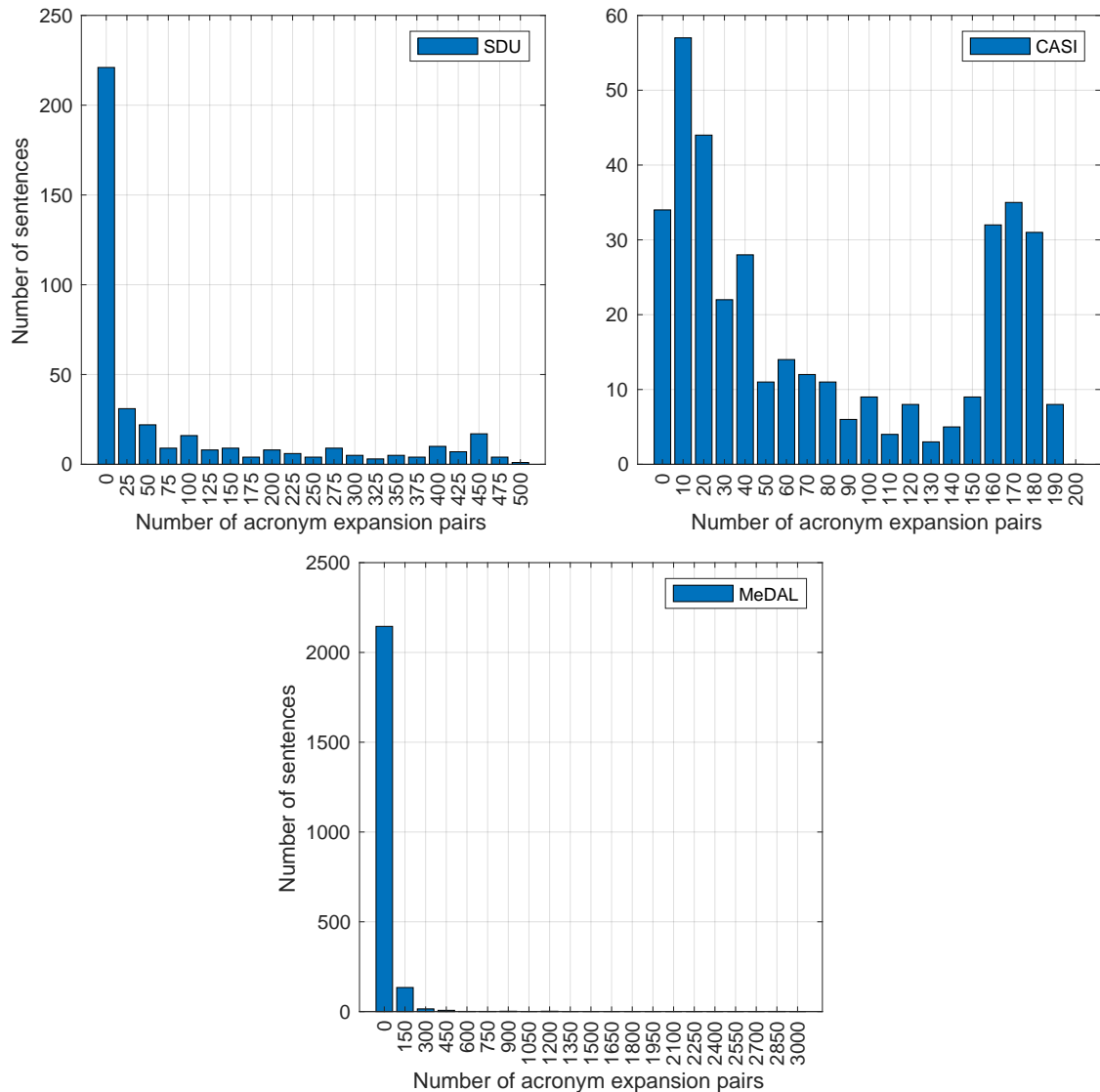


Figure 2: The distribution of samples based on the number of acronym expansion pairs for SDU, CASI, and MeDAL datasets.

C Implementation Details

Our implementation used the pre-trained [SciBERT](#) and [BioBERT](#) model architectures. We conducted out experiments on 1 RTX 3090 graphics cards with 24 GB memory and CUDA 11.4. Our implementation is based on PyTorch 1.8.2.

D Sample Output Representations

Figure 3 shows sample output representations obtained by the m network-based architecture for the SDU, CASI, and MeDAL datasets by reducing the dimensions using PCA. For the SDU dataset, the acronym *RL* with *reinforcement learning* were used to obtain the positive and respective negative sentences. Similarly, for CASI dataset the

acronym *DM* with *diabetes mellitus* expansion and for MeDAL dataset the acronym *RMS* with *respiratory muscle strength* expansion were used.

	Data subset	No. samples	Ratio
SDU	Training	50,034	80%
	Development	6,189	8%
	Test	6,218	12%
CASI	Training	29,600	80%
	Development	3,700	10%
	Test	3,700	10%
MeDAL	Training	24,000	80%
	Development	3,000	10%
	Test	3,000	10%

Table 3: Dataset Statistics

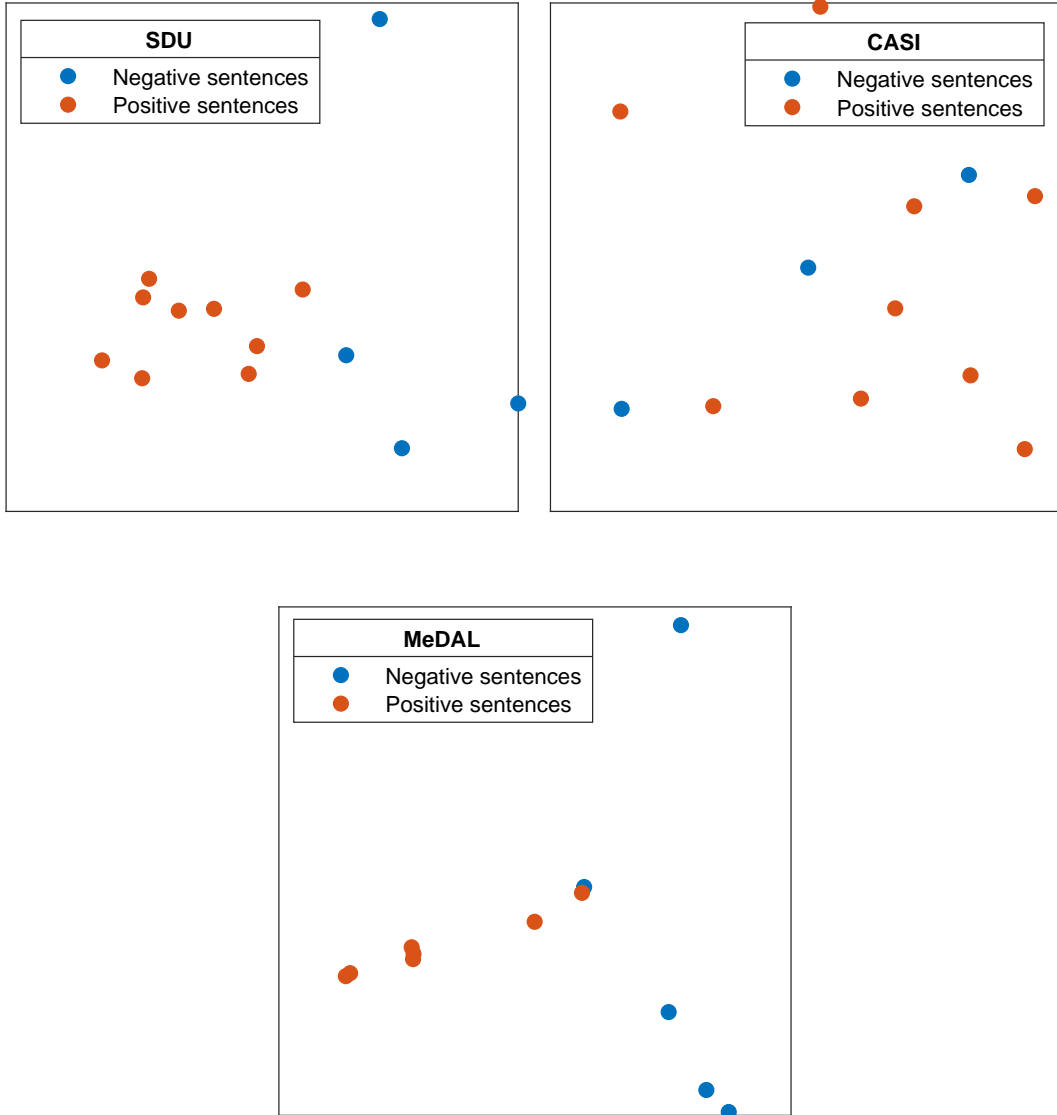


Figure 3: Positive and negative representations obtained by m network-based architecture for the three datasets. For the SDU dataset, the acronym *RL* with *reinforcement learning* were used to obtain the positive and respective negative sentences. Similarly, for CASI dataset the acronym *DM* with *diabetes mellitus* expansion and for MeDAL dataset the acronym *RMS* with *respiratory muscle strength* expansion were used.