Symptom normalization using unsupervised learning and text similarity

Berke Kavak, Arzucan Özgür
Department of Computer Engineering, Bogazici University, Istanbul Turkey
E-mail: berke.kavak@boun.edu.tr, arzucan.ozgur@boun.edu.tr

Abstract

Mapping named entities to their respective IDs in raw texts is an important task, as inaccuracies can significantly affect data consistency and the precision of information retrieval. This is especially important in medical texts, where correct entity identification can have a major impact on diagnostic accuracy and patient care. SympTEMIST is a shared task dedicated to, as the name suggests, text mining of medical symptoms, signs, and findings from texts. In this paper, we present our team BounNLP's participation in subtask 2, which primarily aims to map Spanish symptom mentions to the corresponding SNOMED CT concept IDs. We propose an unsupervised approach for named entity normalization based on clustering and text similarity, using both string similarity and BERT-based contextual word vector representations.

Introduction

In the domain of biomedical entity representations, research has generally centered around the adoption of biomedical word embeddings, as initially discussed by Sung (4). Following the emergence of Word2Vec by Mikolov, Mondal et al. harnessed the biomedical version of Word2Vec embeddings, coupled with a Triplet CNN, to address the Named Entity Normalization (NEN) problem (7). Subsequently, the introduction of state-of-the-art transformer architectures by Devlin et al. in 2019 ushered in a new era of more precise modeling (1).

Within the medical domain, BioBERT gained prominence, serving as the cornerstone for various high-performance models. Sung et al. 2020 (4) introduced BioSyn, a methodology that leverages BioBERT to learn from incomplete synonyms, selecting candidates that maximize the marginal likelihood. Building upon this foundation, Sung et al. 2022 (5) further advanced the field with the introduction of BERN2, a system that seamlessly combines Named Entity Recognition (NER) and Named Entity Normalization (NEN) while also implementing efficiency enhancements to augment the capabilities of BioSyn (5).

The challenge for the NEN task is that word embedding-wise similar mentions can correspond to two different entities. On the other hand, two entities with distant vector representations can refer to the same entity (4). To find the entities, we propose a fast method that covers exact and relaxed dictionary matching, clustering of the same entities, and word similarity techniques. The other work done in the named entity normalization field generally focuses on supervised learning. However, here we deliver a simple approach that is less computationally intensive.

We describe the BounNLP team's method to solve the second task of the SympTEMIST Shared Task (10). The aim of the second task is to correctly identify the SNOMED IDs of the given symptoms in Spanish text. Symptom mentions and their corresponding IDs from a sample text in the data set are provided in Table 1.

Table 1: Symptoms and their SNOMED IDs in a sample text.

mentions	id
masa en el lóbulo tiroideo	237557003
origen tiroideo de la tumoración	237557003
masa encapsulada con respecto al parénquima renal	309088003
masa renal	309088003
masa sólida en polo superior de riñón	309088003

As illustrated in Table 1, the same entity may appear in the same text file multiple times with various mentions. We implement a hybrid approach that maps the mentions to a dictionary that is constructed with the provided gazetteer and the training set. If a named entity mention is not available in the dictionary, a clustering and text similarity-based approach is applied for normalization. The clustering method is offered to improve the accuracy by grouping symptom mentions with similar word embeddings in the same text file so that they are all assigned to the same ID. In addition, we utilize a string similarity method to normalize symptom mentions that remain unnormalized after the clustering step.

Material and Methods

An overview of the proposed system is presented in Figure 1. First, the gazetteer and the training data are used directly to match the entities. The gazetteer is created from the SNOMED database and provided to the shared task participants. Preprocessing was made for enhanced embeddings. These embeddings are prepared using a Spanish BERT model. Employing a clustering approach, we group related mentions within documents. Subsequently, we apply the Jaro Winkler word similarity metric, first within each document, and then normalize the remaining mentions with the most similar terms from the dictionary.

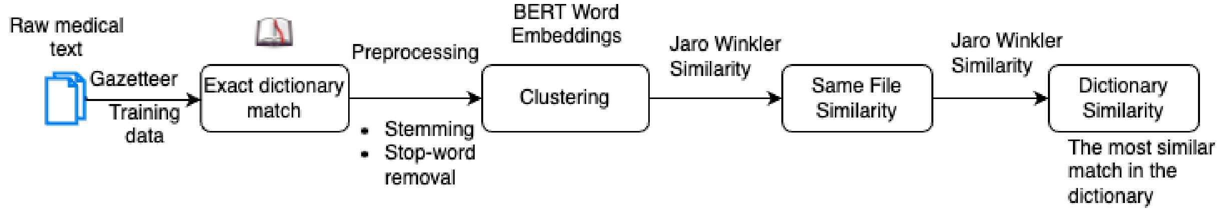

Figure 1: Overview of the proposed approach.

1. Exact Dictionary Match

The training set consists of 304 files and 3484 symptom mentions. In the test set, 3104 mentions will be normalized. First, a dictionary is created with the provided gazetteer and the training set documents. In the first stage, the mentions are directly matched with their dictionary correspondence after being lowercased.

2. Preprocessing

After the exact match, we did stemming and removed Spanish stop words from the symptom mentions using the NLTK library (8) to improve clustering and similarity calculations. The non-digit and non-alpha characters were also removed.

3. Clustering

The provided shared task dataset was separated into distinct files. To enable a comprehensive understanding of the textual content and to identify related entities within each file, we used BERT embeddings to represent the content of the mentions and to capture the semantic relationships and similarities between them. We used the Spanish Pretrained BERT model (9).

The main goal of the clustering approach is to identify symptom mentions that have the same ID in each file. The choice of the clustering algorithm plays a pivotal role in our approach. For this purpose, we adopt the Density-Based Spatial Clustering of Applications with Noise (DBScan) (2). DBScan is particularly advantageous in our context due to its ability to determine the number of clusters automatically, thereby eliminating the need for prior assumptions about the cluster count. To determine the parameters for clustering, we used the silhouette score and purity as optimization metrics. After optimization, the epsilon parameter is set to 0.07. Symptom mentions in a file are represented with their BERT embeddings and clustered using the DBScan algorithm. A cluster may contain symptom mentions that have already been matched to their corresponding IDs using the exact dictionary matching approach, as well as symptom mentions that could not be normalized using exact matching. For such mentions, this step assigns the most frequent ID in the cluster obtained using exact matching.

4. Same File Similarity

After completing the clustering process, we turned our attention to the mentions that had not yet been normalized. Within each file, we performed word similarity analysis. This intra-file similarity analysis was performed because medical texts often contain similar mentions with the same ID within the same document. A non-normalized entity mention is assigned to the ID of the most similar symptom mention in the same file if the similarity is above a predefined threshold. Using the dataset, a comprehensive comparative analysis of similarity methods such as Jaccard, Levenshtein, and Jaro Winkler (3) was performed. Remarkably, the Jaro Winkler similarity metric proved to be the best and consistently produced better results compared to the others. To optimize the accuracy of our system, we fine-tuned the Jaro Winkler similarity metric by experimenting with different thresholds on the validation set, which was obtained by randomly selecting 20% of the training set.

5. Dictionary Similarity

We expanded our approach to inter-file analysis, applying the same similarity technique across the entire dictionary. This step allowed us to successfully retrieve the remaining IDs for non-normalized mentions, further enhancing the overall normalization process. To achieve this, we searched the entire dictionary for the most similar symptom, identified by the highest Jaro Winkler similarity score. Subsequently, we tagged the mention with the same ID as the most closely matching term. This approach highly increased the accuracy of our normalization pipeline.

Results and Discussion

We separated 20% of the training files as the validation set to tune our system. The top-1 accuracies over the validation set are shown in Table 2. Exact dictionary matching achieved 40.85% accuracy, DBScan clustering improved it to 41.36%, and applying the same file similarity analysis further increased the accuracy to 41.67%. Finally, when the entire dictionary is considered for similarity analysis, our system achieved 48.98% accuracy. The developed system achieved a similar accuracy (i.e., 47.21%) on the official test set of the shared task.

Table 2: Summary of the results on the validation set.

	Method			
	Exact Dictionary Match	**Clustering**	**Similarity Same File**	**Similarity All Dictionary**
Accuracy	40.85%	41.36%	41.67%	48.98%

Conclusion

We introduced a hybrid approach for named entity normalization that combines dictionary matching, unsupervised learning, and word similarity techniques. In order to identify the entity mentions with the same category (ID) in a file, the proposed approach uses clustering based on the BERT-based vector representations of the entity mentions. Although the developed system achieved promising results, the clustering approach did not lead to a remarkable increase in normalization performance because only a few files in the shared task dataset contain multiple symptom mentions with the same category.

Our system combines unsupervised learning and text similarity, which is a promising method to address the challenges of named-entity normalization in the medical domain. Further research and improvements in this direction have the potential to provide even more robust and accurate results in the future.

Acknowledgments

TUBA-GEBIP Award of the Turkish Science Academy (to A.O.) is gratefully acknowledged.

References

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Bidirectional Encoder Representations from Transformers. arXiv preprint arXiv:1810.04805.
2. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226-231).
3. Winkler, W. E. (2006). A Comparison of String Distance Metrics for Name-Matching Tasks. In Proceedings of the International Workshop on Information Integration on the Web (IIWeb-06), pp. 73-78.
4. Sung, M., Jeon, H., Lee, J., & Kang, J. (2020). Biomedical Entity Representations with Synonym Marginalization. arXiv preprint arXiv:2005.00239v1.
5. BERN2: an advanced neural biomedical named entity recognition and normalization tool. Oxford, Bioinformatics, 2022, 1–3. Available at: https://doi.org/10.1093/bioinformatics/btac598. (Advance Access Publication Date: 2 September 2022).
6. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. Advances in Neural Information Processing Systems, 26, 3111-3119.
7. Mondal, I., Purkayastha, S., Sarkar, S., Goyal, P., Pillai, J.K., Bhattacharyya, A., Gattu, M. (2020). Medical Entity Linking using Triplet Network. arXiv preprint, arXiv:2012.11164v1.
8. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc.
9. Cañete, J., Chaperon, G., Fuentes, R., Ho, J-H., Kang, H., & Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. In PML4DC at ICLR 2020.
10. "Lima-López, S., Farré-Maduell, E., Gasco-Sánchez, L., Rodríguez-Miret, J. and Krallinger, M. (2023) Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In: Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models.