Medical Entity Linking using Triplet Network

[1]**Ishani Mondal** [1]**Sukannya Purkayastha** [1]**Sudeshna Sarkar** [1]**Pawan Goyal**
[2]**Jitesh K Pillai** [2]**Amitava Bhattacharyya** [2]**Mahanandeeshwar Gattu**
[1]Department of Computer Science and Engineering, IIT Kharagpur
[2]Excelra Knowledge Solutions Pvt Ltd, Hyderabad, India

Abstract

Entity linking (or Normalization) is an essential task in text mining that maps the entity mentions in the medical text to standard entities in a given Knowledge Base (KB). This task is of great importance in the medical domain. It can also be used for merging different medical and clinical ontologies. In this paper, we center around the problem of disease linking or normalization. This task is executed in two phases: candidate generation and candidate scoring. In this paper, we present an approach to rank the candidate Knowledge Base entries based on their similarity with disease mention. We make use of the Triplet Network for candidate ranking. While the existing methods have used carefully generated sieves and external resources for candidate generation, we introduce a robust and portable candidate generation scheme that does not make use of the hand-crafted rules. Experimental results on the standard benchmark NCBI disease dataset demonstrate that our system outperforms the prior methods by a significant margin.

1 Introduction

A disease is an abnormal medical condition that poses a negative impact on the organisms and enabling access to disease information is the goal of various information extraction as well as text mining tasks. The task of disease normalization consists of assigning a unique concept identifier to the disease names occurring in the clinical text. However, this task is challenging as the diseases mentioned in the text may display morphological or orthographical variations, may utilize different word orderings or equivalent words. Consider the following examples:

Example 1: "..characteristics of the disorder include a **short trunk and extremities**..." **Source** : (PMID:7874117)
Example 2: "**Renal amyloidosis**, prevented by colchicine, is the most severe complication of FMF ..." **Source** : (PMID:10364520)

In Example 1, the disease mention **short trunk and extremities** should be mapped to a candidate Knowledge Base entry(D006130) containing synonyms like **Growth Disorder**. In Example 2, **Renal amyloidosis** should be assigned to Knowledge Base ID (C538249) which has synonyms such as, **Amyloidosis 8**.

Based on our studies and analysis of the medical literature, it has been observed that the same disease name may occur in multiple variant forms such as. synonyms replacement (e.g.*"lung cancer"*, *"lung carcinoma"*), spelling variation (*"Acetolysis"*, *"acetolisis"*), a short description modifier precedes the disease name (e.g. *"massive heart attack"*), different word orderings (eg. *"alpha-galactosidase deficiency"*, *"deficiency of alpha-galactosidase"*).

In this paper, we have formulated the task of learning mention-candidate pair similarity using Triplet Networks (Hoffer and Ailon, 2015). Furthermore, we have explored in-domain word[1] and subword embeddings (Bojanowski et al., 2017) as input representations. We find that sub-word information boosts up the performance due to gained information for out-of-vocabulary terms and word compositionality of the disease mentions.

The primary contributions of this paper are three-fold: 1) By identifying positive and negative candidates concerning a disease mention, we optimize the Triplet Network with a loss function that influences the relative distance constraint 2) We have explored the capability of in-

[1]http://evexdb.org/pmresources/vec-space-models/

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 95–100
Minneapolis, Minnesota, June 7, 2019.

Dataset	Abstracts	Total	Unique
Training set	692	5932	1538
Test set	100	960	427
Total	792	6892	1965

Table 1: NCBI Disease Corpus Statistics

domain sub-word level information[2] in solving the task of disease normalization. 3) Unlike existing systems (D'Souza and Ng, 2015), (Li et al., 2017), we present a robust and portable candidate generation approach without making use of external resources or hand-engineered sieves to deal with morphological variations. Our system achieves state-of-the-art performance on NCBI disease dataset (Dogan et al., 2014)

2 Dataset

The NCBI disease corpus (Dogan et al., 2014) contains 792 Pubmed abstracts with disorder concepts manually annotated. In this dataset, disorder mentions in each abstract are manually annotated with the identifier of the concept in the reference ontology to which it refers. It uses MEDIC lexicon (Davis et al., 2012) as the reference ontology. (See Table 1 for dataset statistics)

3 Methodology

The dataset consists of a certain number of abbreviations, in order to identify these cases, we have considered the mentions composed of all uppercase letters as abbreviations. We find the disease mentions immediately preceding the abbreviated terms and substitute all the occurrences of the abbreviated words in that document with the corresponding expanded disease mentions. Our system primarily consists of two modules: 1) **Candidate generation**: (See section 3.1) Generate potential candidates from the Knowledge Base corresponding to a disease mention. 2) **Candidate ranking**: (See section 3.2) Rank those potential candidates corresponding to a disease mention.

3.1 Candidate generation

In this section, we discuss the algorithm which generates the potential candidates to which the disease mentions might be referring. In this study, the Knowledge Base entries were sampled from the entire MEDIC Lexicon, but not limited to only annotations in the NCBI Disease Corpus.

For a given disease mention, the candidate generation algorithm generates candidates from the Knowledge Base entries. Suppose, the Knowledge Base consists of k entries, each having a certain number of synonyms. Each multi-word synonym represented by the sum of its word embeddings. For a given mention m consisting of l words represented by $\{m_1, m_2, \ldots, m_l\}$, we represent m as the sum of its word embeddings. The steps for the candidate generation algorithm are as follows:

- **Step 1**: Candidate Set 1, $\{C_1\}$: Calculate the cosine similarity between vector representation of each synonym (candidate) of the KBIDs and the mention. Identify the top k_1 ids whose candidates have cosine similarity greater than or equal to threshold t_1.

- **Step 2**: Candidate Set 2, $\{C_2\}$: Calculate the Jaccard overlap of the mention and the candidates of each KBID. Identify the top k_2 ids having Jaccard overlap score greater than or equal to threshold t_2.

Note: min ($\mid C_1 \mid$,$\mid C_2 \mid$) $\leq \mid C_1 \cup C_2 \mid \leq (k_1 + k_2)$
In our experiments, we choose $t_1 = 0.7$, $t_2 = 0.1$, $k_1 = 3$ and $k_2 = 7$.

We provide examples of candidates generated from our proposed algorithm below.

Mention: "bacteremic infections due to Neisseria"
Candidate Set 1, $\{C_1\}$ = {"bacterial neisseria infections"}
Candidate Set 2, $\{C_2\}$ = {"bacterial neisseria infections" , "DNA-virus infections" , "Screw-Worm Infections" }

3.2 Candidate Ranking

Assume that there are n candidates represented by $\{c_1, c_2, \ldots, c_n\}$ for an entity mention m, we use a Triplet Network which has proven to perform well in many Computer Vision (Hoffer and Ailon, 2015) as well as Natural Language Processing tasks (Clark and Manning, 2016) . As such given a triplet, the idea is to leverage the notion of reducing the distance between the mention and its positive candidate while increasing the distance between the mention and its negative candidate.

[2]https://github.com/ncbi-nlp/BioSentVec.git

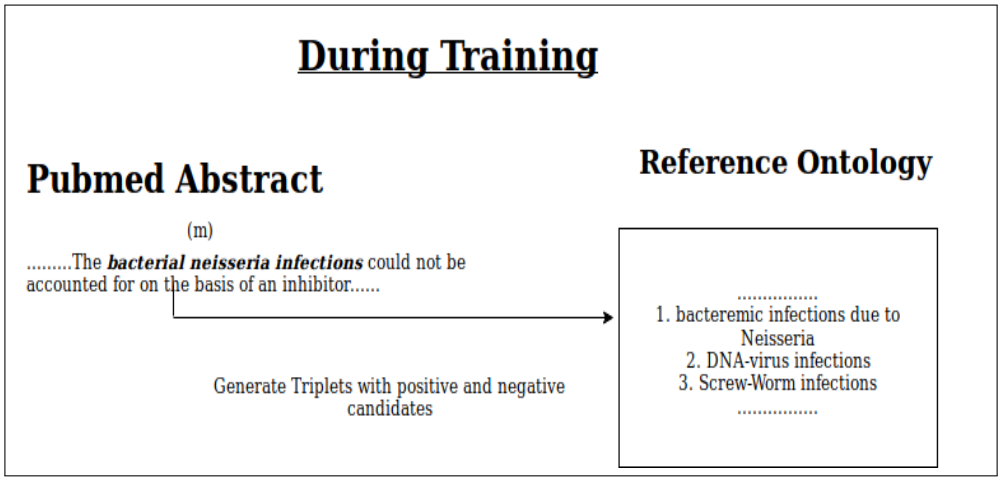

Figure 1: Pictorial Representation of the training Data Generation Process

3.2.1 Triplet data generation

In order to learn better semantic representations between a disease mention and its corresponding candidates, we have generated training data in the form of triplets consisting of disease mention m, positive candidate q_p, negative candidate q_n. The triplet is represented as (q_p, m, q_{n_i}) where $i \in \{1 \ldots [|k_1 \cup k_2| - 1]\}$.

An example of the triplet data is given below:

Disease Mention: "bacteremic infections due to Neisseria"
Positive Candidate: "bacterial neisseria infections"
Negative Candidates: "DNA-virus infections", "Screw-Worm Infections".
The triplets are as follows:

- ("bacterial neisseria infections", "bacteremic infections due to Neisseria", "DNA-virus infections")
- ("bacterial neisseria infections", "bacteremic infections due to Neisseria", "Screw-Worm Infections")

3.2.2 Model Architecture

The Triplet Network architecture as proposed by (Hoffer and Ailon, 2015) has been adopted for the task of entity normalization. To train the model, each triplet consisting of mentions and its candidates are fed into the parameter-shared network ($Conv$), as a sequence of word embeddings. For a triplet, (q_p, m, q_{n_i}) the layer outputs their representations $Conv(q_p)$, $Conv(m)$ and $Conv(q_{n_i})$ respectively. Our objective is to make the representations of m and q_p closer than the representations of m and q_{n_i}. Thus the next layer uses a distance function, denoted by dis, to compute the distances as follows:

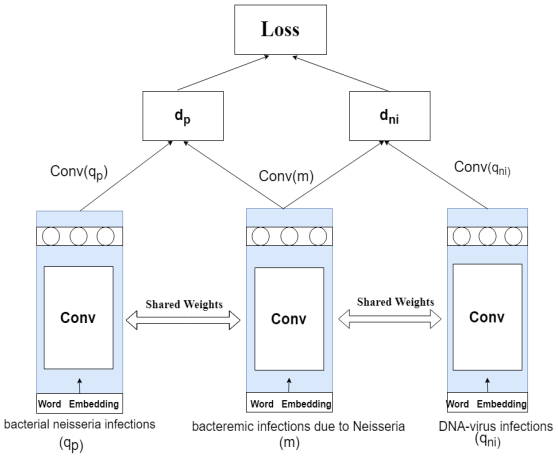

Figure 2: The word and sub-word embeddings of triplet ('bacterial neisseria infections', 'bacteremic infections due to Neisseria', 'Screw-Worm Infections') are fed as batches into the Triplet Network.

$$d_p = dis(Conv(m), Conv(q_p))$$

$$d_{n_i} = dis(Conv(m), Conv(q_{n_i}))$$

Here d_p specifies the distance between target disease mention m and q_p while d_{n_i} specifies the distance between target disease mention m and q_{n_i}. The triplet loss function (L) used for achieving this goal has been formulated as follows:

$$L = \max(d_p - d_{n_i} + \alpha, 0)$$

Another variable α, a hyperparameter is added to the loss equation which defines how far away

the dissimilarities should be. Thereafter, by using this loss function, we calculate the gradients and update the parameters of the network based on these gradient values. For training the network, we take mention m and randomly sample q_p and q_{n_i} and compute their loss function and update their gradients.

We use 200-dimensional word2vec (Mikolov et al., 2013) embeddings trained on Wikipedia and Pubmed PMC-Corpus (Pyysalo et al., 2013) as input to $Conv$. To deal with the huge number of out-of- vocabulary terms in the medical domain, we have used the $fastText$ based sub-word embeddings (Galea et al., 2018). $fastText$ (Bojanowski et al., 2017) has been applied on PubMed and MIMIC-III (Johnson et al., 2016) to generate 200- dimensional word embeddings, the window size being 20, learning rate 0.05, sampling threshold 1e-4, and negative examples 10 (yijia zhang et al., 2018).

3.2.3 Training Details

$Conv$ is composed of one convolutional and max-pooling layer. ReLU non-linearity (Maas, 2013) is applied between two consecutive layers. The final embedding of $Conv$ is a fixed-length(128) vector. For dis and the loss function we use the L2 distance (Danielsson, 1980). The triplet loss has been applied. For training we use Adam Optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.001. Training has been done for 50 epochs, and early stopping has been employed on the basis of the accuracy of the validation set. After hyperparameter tuning, several experiments have been performed, and the results on the best hyperparameter settings have been reported.

3.2.4 Evaluation

After the model has been trained, we evaluate the rank of each of the disease mentions in the test set. For each of the disease mention m in the test set, we run the candidate generation algorithm to find out the maximum cosine similar candidates for the potential KnowledgeBaseIDs. The positive candidate is labelled as 1 while the rest has been labelled as 0. During the process of evaluation, we calculate the similarity score between the disease mention and its candidates. The similarity scores are then sorted in descending manner in order to rank the candidates based on its similarity. We choose the candidate with the maximum similarity score for each of the disease mentions.

Model Name	Accuracy
(D'Souza and Ng, 2015)	84.65
(Li et al., 2017)	86.10
Triplet CNN + static word2vec	86.09
Triplet CNN + dynamic word2vec	87.85
Triplet CNN + subword	89.65
Triplet CNN + subword + abb	**90.01**

Table 2: The table shows the accuracy of our system in comparison with the baseline systems.

We choose the evaluation measure as accuracy. Since, the highest similar candidate is of primary interest in the task of entity linking, so we choose the top-K (Where $K = 1$).

TP = It signifies that the highest ranked candidate for disease mention m is the actual referent KnowledgebaseID.

FP = It signifies that the highest ranked candidate for disease mention m is not the actual referent KnowledgebaseID.

$$Accuracy = \frac{TP}{TP + FP}$$

4 Results

We report accuracy for our system in finding the correct Knowledge Base ID corresponding to a disease mention in the text. **Table 2** shows that in comparison with the existing baseline systems, **Subword information** as input to the Triplet Network and abbreviation expansion from the document context (Triplet CNN+subword+abb) performs the best. From the feature ablation, it is clear that the in-domain word embeddings((Triplet CNN + dynamic word2vec) and (Triplet CNN + static word2vec)) are essential for capturing better semantic representations.

5 Analysis

In this section, we throw some light on both the merits and demerits of the proposed system with respect to the baseline models.

5.1 Merit Analysis

We compare our results with other rule-based and neural network based methods known to perform well on this standard dataset. To gain more insights into our proposed model, in particular, the importance of the domain-specific word and subword representation to capture the semantic and

syntactic similarity using Triplet Network, we select some examples from the labeled test set. In figure 2, two different cases have been shown which demonstrate the performance gap between our and the existing baseline systems.

In Example 1, the disease mention *"inherited neurodegeneration"* was not mapped with *"heredodegenerative disorders"* (D020271) by the existing methods, because of their incapability to capture the semantic similarity. In contrast to this, our system obtains additional semantic and syntactic information from the domain-specific subword embeddings and thereby maps to the correct concept ID.

In Example 2, the abbreviation *"AS"* is polysemous in nature as it can either be mapped to the concepts like *"Angelman Syndrome"* (PMID : 9585605) and *"Ankolysing Spondylitis"* (PMID : 9336417). Due to the lack of contextual information in the existing models, they were not able to handle the polysemous nature of the abbreviations; but abbreviation expansion from the document level context in our system handles this scenario.

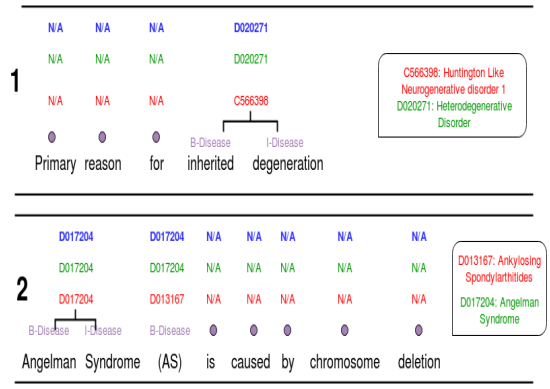

Figure 3: The NER tags as input are shown in purple, the gold standard conceptID is shown in green, the predictions from the baseline systems are shown in red, whereas the prediction from the proposed system is shown in blue.

5.2 Demerit Analysis

The error types incurred by the proposed system have been explained in detail as follows:

1) **Ambiguous distribution of importance to the disease name**: The system fails to understand which part of disease mention to provide more attention while performing normalization. Suppose the disease mention is **"colorectal adenoma"**, during normalization, the system mistakenly normalizes the disease to the concept ID predominated by **"colorectal"**. Automatic identification of such semantic attention is challenging and deserves a significant spot in the future research.

2) **Incorrect mapping of certain ambiguous disease names**: Suppose the disease mention dysmorphic features in "..loss of MAGEL2 may be critical to abnormalities in brain development and **dysmorphic features** in individuals with PWS.." (PMID: 10915770) has been mapped to D057215 whereas the same disease mention in "..She had minor **dysmorphic features** consistent with those of.." (PMID: 8071957) has been assigned to D000013. Since, in these two examples, the disease mention in these two examples have been assigned as "diseaseClass" and "Modifier" features respectively. It happens due to different NER features of the mention annotated in the dataset. But incorporating this NER feature in our proposed model unnecessarily generates huge number of false positives.

6 Conclusion

In this paper, we have formulated the task of entity linking as a candidate ranking approach. Using a Triplet Network, we learn high-quality representations of candidates, tailored to reveal relative distances between the disease mention and its positive and negative candidates. Furthermore, we take a step towards eliminating the need to generate candidates based on hand-crafted rules and external knowledge resources. Though our method outperforms the existing systems by a strong margin, there is a scope for improvement in terms of attention-based disease similarity (viz, "Neisseric infections" imply the importance of "Neisseric" during its similarity computation with the "bacterial neisseria infections"). An intriguing course for future work is to further explore the robustness and scalability of this approach to other clinical datasets for entity normalization.

Acknowledgments

This work has been supported by the project "Effective Drug Repurposing through literature and patent mining, data integration and development of systems pharmacology platform" sponsored by MHRD, India and Excelra Knowledge Solutions, Hyderabad. Besides, the authors would like to thank the anonymous reviewers for their valuable comments and feedback.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262. Association for Computational Linguistics.

Per-Erik Danielsson. 1980. Euclidean distance mapping. *Computer Graphics and Image Processing*, 14(3):227 – 248.

Allan Peter Davis, Thomas C. Wiegers, Michael C. Rosenstein, and Carolyn J. Mattingly. 2012. Medic: a practical disease vocabulary used at the comparative toxicogenomics database. *Database (Oxford)*, 2012:bar065–bar065. 22434833[pmid].

Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *J Biomed Inform*, 47:1–10. 24393765[pmid].

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 297–302.

Dieter Galea, Ivan Laponogov, and Kirill A. Veselkov. 2018. Sub-word information in pre-trained biomedical word representations: evaluation and hyperparameter optimization. In *BioNLP*.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *SIMBAD*.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035 EP –. Data Descriptor.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Haodi Li, Qingcai Chen, Buzhou Tang, Xiaolong Wang, Hua Xu, Baohua Wang, and Dong Huang. 2017. Cnn-based ranking for biomedical entity normalization. *BMC Bioinformatics*, 18(Suppl 11):385–385. 28984180[pmid].

Andrew L. Maas. 2013. Rectifier nonlinearities improve neural network acoustic models.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Sampo Pyysalo, F Ginter, Hans Moen, T Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. *Proceedings of Languages in Biology and Medicine*.

yijia zhang, qingyu chen, zhihao yang, hongfei lin, and Zhiyong Lu. 2018. BioWordVec: Improving Biomedical Word Embeddings with Subword Information and MeSH Ontology.