ULisboa: Recognition and Normalization of Medical Concepts

André Leal[+], Bruno Martins[*], and Francisco M. Couto[+]
[+]LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal.
[*]INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Portugal
aleal@lasige.di.fc.ul.pt , bruno.g.martins@ist.ul.pt , fcouto@di.fc.ul.pt

Abstract

This paper describes a system developed for the disorder identification subtask within task 14 of SemEval 2015. The developed system is based on a chain of two modules, one for recognition and another for normalization. The recognition module is based on an adapted version of the Stanford NER system to train CRF models in order to recognize disorder mentions. CRF models were build based on a novel encoding of entity spans as token classifications to also consider non-continuous entities, along with a rich set of features based on (i) domain lexicons and (ii) Brown clusters inferred from a large collection of clinical texts. For disorder normalization, we (i) generated a non ambiguous dictionary of abbreviations from the labelled files, using it together with (ii) an heuristic method based on similarity search and (iii) a comparison method based on the information content of each disorder. The system achieved an F-measure of 0.740 (the second best), with a precision of 0.779, a recall of 0.705.

1 Introduction

Clinical notes are an important source of information recorded by medical professionals. However, this information, when available, is not easily accessible within automated procedures. Clinical notes are inherently complex, due to their lack of structure (i.e., narrative language) and due to the need for contextual interpretation. To address this complexity, text mining approaches represent an effective solution to assist the users in retrieving and extracting the required information.

This paper presents a text mining system for processing clinical text, that we developed for SemEval based on a pipeline with two modules, one for entity recognition and another for normalization.

The entity recognition module is based on the Stanford NER tool (Finkel et al., 2005), and it uses CRF models trained on annotated biomedical notes. The module tags the text according to an SBIEON encoding of entities as token classes, supporting the recognition of non-continuous entities (Leal et al., 2014). We relied on features based on Brown clusters and domain specific lexicons. Thus, this approach combines both supervised (Stanford NER) and unsupervised methods (Brown Clusters).

For practical applications, entity recognition is incomplete without performing normalization, i.e. without mapping each entity to an identifier (CUI) in a controlled vocabulary like SNOMED CT (Cornet and Keizer, 2008), that defines its semantic meaning. One of the main challenges in this task consists in resolving the ambiguous cases, where the same entity can have distinct semantic meanings (i.e., mapped to distinct CUIs) depending on the context.

Our normalization module relies on the following components: (i) a procedure for the automatic generation of auxiliary dictionaries from the labelled training data (e.g, abbreviations) and from SNOMED CT, to be used as mapping dictionaries, (ii) an heuristic for similarity search, and (iii) an information content measure for each concept.

Our system is an extension of the one used in the 2014 edition of SemEval (Leal et al., 2014). Both systems used the same approach for entity recognition but, in terms of the normalization component, the system from 2014 was entirely based on a lexical similarity approach using NGram, Levenstein

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 829–834,
Denver, Colorado, June 4-5, 2015.

and JaroWinkler distances. The current system is instead based on a pipeline were the information content was also incorporated. Besides SNOMED CT, the current system also integrated dictionaries automatically generated from the training data.

2 The SemEval Task

Task 14 of SemEval 2015 was composed of two subtasks: recognition and normalization of medical concepts (subtask 1) and disorder slot filling (subtask 2). We only participated in subtask 1.

The recognition part of subtask 1 consisted on performing the recognition of medical concepts, who belong to the UMLS semantic group *disorders*, within unstructured clinical notes. The disorders group of UMLS corresponds to concepts defined within SNOMED CT (Cornet and Keizer, 2008). Recognized entities can be continuous, non-continuous or even overlapped in the text.

The normalization part consisted on the mapping of an unique UMLS CUI (Concept Unique Identifier) to each previously recognized entity, or none at all (CUI-Less) for the cases where there is no suitable CUI for the recognized entity within the SNOMED CT database. Ambiguous entities represent the main challenge of this task, since identifying the correct CUI depends on their context.

Task 14 evaluated the recognition and normalization parts as one single task, by measuring the final system's precision, recall and F-measure. The evaluation could also be performed in a strict or relaxed way. In strict evaluation, a predicted mention is considered a true positive if the predicted span is exactly the same as the gold-standard. On the relaxed evaluation, the predicted spans only need to overlap the gold-standard spans to be considered a true positive. On both evaluation methods the CUI must be correctly identified to be considered a true positive. Thus, even with a perfect recognition system, it is possible to achieve low results on the task, depending on the normalization performance

3 Datasets

Similarly to the last edition of the competition (Zhang et al., 2014), two sets of labelled data were given to the participants, which were separated into two categories (training and development). They were used for training and testing of our system, respectively. Unlabelled clinical notes from the MIMIC corpus were also provided. Later, an unlabelled test set was released to evaluate the final system. Unlabelled clinical notes consisted on plain text without any additional information, while labelled clinical notes consist on plain text together with a list of disorder mentions contained on them. Table 1 summarizes each dataset.

	Train	Devel	Test	Unlabelled
Notes	298	133	100	404k
Words	182k	154k	8k	123M
Disorder Mentions	11.5k	8k	-	-
CUI-ied	8k (88%)	6k (76%)	-	-
CUI-less	3.5k (12%)	2k (24%)	-	-

Table 1: Statistical characterization of the datasets.

4 Entity Recognition

We applied the same type of approach used in our system from last year (Leal et al., 2014) for entity recognition. The Stanford NER software (Finkel et al., 2005) was used to train Conditional Random Fields (CRF) models using labelled data as input.

All input text had to be tokenized and encoded according to a named entity recognition scheme that encodes entities as token classifications. To be able to recognize non-continuous entities, an SBIEON (Leal et al., 2014) encoding was used. Besides the tags defined in the SBIEO encoding (Ratinov and Roth, 2009), a new tag **N** was added to identify words that do not belong to the entity but are inside the continuous span that contains the recognized entity. The remain tags are used to identify **S**ingle entities, the **B**egin, **I**nside and **E**nd token of a non-single token entity, and the **O**ther tag for words which are neither entities nor related to them. For overlapped entities we did not develop any approach, i.e. we only recognize the first entity in an overlapping group of entities. Thus, handling overlapping entities remains an open issue in our system.

4.1 Recognition Features

We generated 2nd-order CRF models by using, as training data, the labelled notes together with a rich set of features. In 2nd-order models, the features

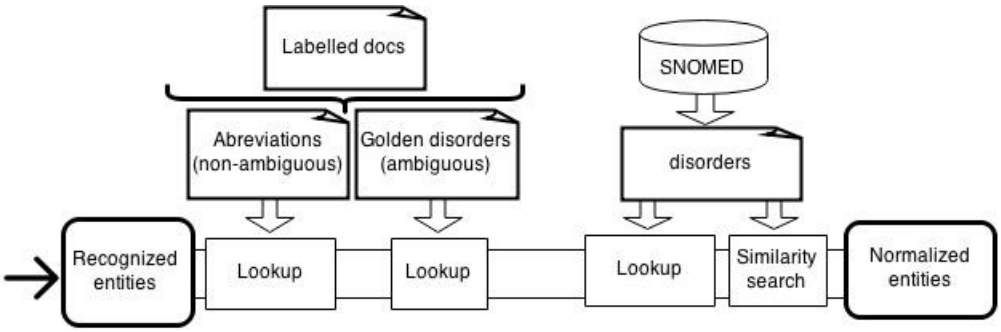

Figure 1: Overview on the normalization approach.

are computed from representations composed by the current class and the two previous/next classes.

Training Data: Two different sets of data were employed: one with notes belonging to the training set only, and another with notes from both the training and devel sets.

Brown Clusters: We inferred word representations in the form of Brown clusters (Brown et al., 1992) from all data that was made available, i.e. from MIMIC, train and devel. According to (Turian et al., 2009), this technique reduces the data sparsity, generating lower-dimensional representations of the word vocabulary, and therefore increasing the accuracy. Each word cluster contains a group of words, and clusters are formed by maximizing the mutual information of bi-grams, according to a class-based language model. We used a total of 404k documents, containing an approximate total of 123M tokens, to infer 100 different clusters using the implementation provided by (Turian et al., 2010). The number of clusters was chosen through a separate set of experiments as the one that maximized the F-measure.

Encoding: The aforementioned SBIEON encoding was employed in all recognition models.

Features: The CRF models rely on a set of features that includes (i) word tokens within a window of size 2, (ii) token shape (upper-cased, numeric, etc), (iii) token position in a sentence and (iv) token prefixes and suffixes. This basic set of features was also extended with features based on Brown clusters, and domain-specific lexicons.

Domain-specific lexicon: We built lexicons for the medical domain that include (i) SNOMED CT disorders, (ii) drugs and diseases from DBPedia and (iii) a list of disorders from the labelled data.

5 Normalization

Each recognized entity needs to be normalized, if possible, with a unique identifier (CUI) from an existing controlled vocabulary. This way, a semantic meaning is associated to each entity. Since ambiguous entities can have multiple identifiers depending on the context, one of the main challenges in this task consists in the disambiguation of these cases.

To address this challenge, we developed a pipeline framework (Figure 1) composed of several modules. First, a recognized entity will be looked up in an abbreviation dictionary. If it is unambiguously present there, then the associated CUI is assigned, otherwise the entity moves on to the next module (i.e. lookup on the golden dictionary). The CUI-less tag is assigned to the entity if no suitable CUI is found at the end of this process, or if the most similar SNOMED CT candidate found is not a disorder.

5.1 Resources

Abbreviation dictionary: This dictionary contains the small (up to 4 letters) upper-cased non-ambiguous concept descriptors found in the labelled data. For instance, the entity *ASD* is an abbreviation of *atrial septal defect* with the CUI *C0018817*. Since this descriptor is unique in SNOMED CT, it is considered non-ambiguous.

Golden disorders dictionary: All entity spans (ambiguous included) retrieved from the labelled notes are used to form this dictionary. This dictionary is thus composed by all concept descriptors which were dropped by the abbreviation dictionary, for their length or because they were ambiguous.

SNOMED CT dictionary: All concepts from SNOMED CT are included.

5.2 Methods

Similarity Search: This module was implemented using a Lucene index (MacCandless et al., 2010). NGram (Kondrak, 2005) and Levenshtein distances were used to retrieve the best SNOMED CT candidates. An extended Levenshtein distance, based on a best-token-match approach, was developed. This distance gives the similarity between a *target* (recognized entity descriptor) and a *candidate* descriptor (SNOMED CT concept), regardless of their token's orders. First, both *target* and *candidate* descriptors are split into tokens. For each *target*'s token, we compute the Levenshtein distance with all *candidate* tokens, and we finally pick the token corresponding to the minimum value. Each token in the candidate can only be compared to a single token in the target. The distance is represented by the following formula:

$$S_{dt} = \text{SplitTokens}(d_t)$$

$$S_{dc} = \text{SplitTokens}(d_c)$$

$$\text{Sim}(d_t, d_c) = \begin{cases} -1, & \text{if} |S_{dt}| > |S_{dc}| \\ \dfrac{\sum_{w_{dt} \in S_{dt}} \text{BestMatch}(w_{dt}, S_{dc})}{|S_{dt}|}, & \text{otherwise} \end{cases}$$

In the formula, we have that

$$\text{BestMatch}(w_{dt}, S_{dc}) = \text{Min}\{\text{LevDist}(w_{dt}, w_{dc}) : w_{dc} \in S_d c\}$$

In the previous expressions, d_t is the *target* and d_c the *candidate* descriptor. SplitTokens is the function responsible for splitting the descriptor into tokens. BestMatch returns the minimum Levenshtein distance between the token w_{dt} and all available tokens in S_{dc}. The token in S_{dc} which minimizes the Levenshtein distance is removed from the list for posterior iterations against the remain tokens in S_t.

Information Content (IC): The Information Content (IC) was calculated for each disorder entity using the UMLS-Similarity (McInnes et al., 2009) software implementation. This measure enabled us to disambiguate entities by choosing, from the list of candidates, the ones with the lowest IC. This assumes that more general concepts have a higher probability to appear on a text. The intrinsic method by (Sánchez et al., 2012) was chosen to calculate the IC of each concept, using the following formula where leaves(c) represents the number of leaves of c, subsumers(c) represents the number of parents of c, and max.leaves is the number of nodes which are leaves in the SNOMED CT taxonomy:

$$\text{IC}(c) = -\log\left(\frac{\frac{|\text{leaves}(c)|}{|\text{subsumers}(c)|} - 1}{\text{max.leaves} + 1}\right)$$

5.3 Approach

We implemented a lookup method in each dictionary. If the entity was found, then the associated identifier was immediately assigned. Ambiguous cases were resolved using the information content, choosing the concept with the lowest IC value. For descriptions not found in the considered dictionaries, we used Lucene to retrieve the top 300 most similar candidates from SNOMED CT and, for each candidate, we applied the following formula to obtain the final similarity measure:

$$\begin{aligned}\text{Sim}(d_c, d_t) = 0.15 * \text{Lev}(d_c, d_t) + 0.15 * \text{NGram}(d_c, d_t) \\ + 0.7 * \text{LevExt}(d_c, d_t)\end{aligned}$$

In the previous expression, Sim represents the similarity between the target d_t and candidate d_c descriptor. Lev, NGram and LevExt represent the Levenshtein, NGram5 and Extended Levenshtein distance, respectively. The constant values were chosen according to a separate empirical evaluation using the devel dataset, although in future work we intent to use systematic approaches based on learning to rank. For each CUI associated to the chosen candidate descriptor (higher similarity with target descriptor), the one with the lowest IC was chosen.

6 Evaluation Experiments

Three runs were submitted to the SemEval 2015 competition:

Run 1: A 2nd-order CRF model was trained using the SBIEON encoding, and a rich set of features that includes the domain lexicons and 100 Brown clusters. For training, we only used notes from the training set. For assigning a UMLS identifier to each entity, we used the framework that was previously described.

Run	Strict Evaluation			Relaxed Evaluation		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
1	0.748	0.676	0.710	0.782	0.706	0.742
2	0.749	0.681	0.713	0.780	0.709	0.743
3	**0.779**	**0.705**	**0.740**	**0.806**	**0.729**	**0.765**
Best System SemEval 2015	0.783	0.732	0.757	0.815	0.762	0.788

Table 2: The official results for Task 1 of the SemEval 2015 challenge on clinical NLP.

Run 2: This run is identical to **Run 1** with the exception of the domain lexicon features that were not included. Normalization followed the same strategy as in **Run 1**.

Run 3: Identical to **Run 1**, with the exception that both train and devel documents were used as training data, resulting in the addition of 133 notes to the training set.

7 Results and Discussion

We present our official results in Table 2, highlighting our best results in comparison to those of the best participating system in the competition.

Our best run achieved the second best F-measure in the competition, with an F-measure of 0.740 in the strict evaluation and 0.765 in the relaxed evaluation. As previously said, the predicted mention can only be correct if and only if the mapped CUI is correct.

One of the first things to notice when comparing the runs is the difference on the results between the third run and the others. As expected, the addition of 133 notes (devel set) to the training data produced a better recognition model, thus improving the global performance of the system.

The addition of domain lexicon features to the recognition model resulted in a lower precision on the strict evaluation. On the relaxed evaluation a small improvement was achieved.

The small difference between the strict and relaxed evaluation modes can be associated to a really precise recognition model or, more likely, with the normalization pipeline having trouble in normalizing the concepts when they are not fully recognized. For example, if an entity E was only partially recognized, then it will be harder to normalize it.

In what concerns normalization, all runs were produced using the same pipeline and with the same features. Since our approach for the recognition task is similar to the one used in the SemEval 2014 edition, and since a significant improvement in the overall performance was obtained, we can conclude that our recent developments in the normalization part of the system were particularly effective.

8 Conclusions and Future Work

This paper describes our participation in Task 14 of the SemEval 2015 competition. Although this task was divided into two subtasks, our work only addressed on the recognition and normalization of entity disorders on clinical notes.

For the recognition part, we used a similar approach to the one followed in the 2014 edition of SemEval. Specifically, a 2nd-order CRF model was generated using the Stanford NER software, considering different sets of features. All models used the SBIEON encoding (Leal et al., 2014) to support the recognition of non-continuous entities. Overlapped entities continue to be an open issue.

For the normalization part, we developed a pipeline that takes advantage of the existing labelled data to generate and explore auxiliary dictionaries (e.g., an abbreviation dictionary). For the recognized entities that do not match to any dictionary, we employ a similarity search based on Lucene's implementation of Levenshtein and NGram distances. An extension of the Levenshtein distance was developed to compare descriptors independently of the order of their words. Ambiguous cases were resolved by choosing the concepts with the lowest information content, which was calculated using the approach proposed by (Sánchez et al., 2012);

As expected, results show that a more comprehensive training set enables the generation of better recognition models, maintaining the same set of features. We also saw that the addition of a domain lexicon increased the precision, although not

significantly and with almost no impact on the F-measure. Our normalization framework was likely the main reason for the large improvement in our results, when comparing to the results from SemEval 2014.

In our opinion, the evaluation method followed in this year's competition is good for evaluating the system as a whole, but on the other hand it also limits the evaluation of the two tasks separately, which we believe would bring some advantages while developing the system and when comparing results.

For future work, we intend to evaluate both tasks individually, to better understand which components are performing well, and which ones need to be improved. In the normalization task, we intend to improve the framework that was presented, exploring semantic similarity based on ontology relations (Couto et al., 2006). By assuming that concepts within the same text are semantically related to each other, we intend also to disambiguate entities based on their semantic similarity towards all other previously normalized entities (Lamurias et al., 2015).

To improve the module related to similarity search for disambiguation, we also intend to develop a learning to rank approach similar to the one presented by (Leaman et al., 2013).

Acknowledgments

The authors would like to thank Fundação para a Ciência e Tecnologia (FCT) for the financial support of LASIGE (UID/CEC/00408/2013) and INESC-ID (UID/CEC/50021/2013).

References

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based N-gram Models of Natural Language. *Computational linguistics*, 18(4):467–479.

Ronald Cornet and Nicolette de Keizer. 2008. Forty years of SNOMED: A literature review. *BMC Medical Informatics and Decision Making*, 8(Suppl 1:S2):1–6.

Francisco M Couto, Mário J Silva, and Pedro M Coutinho. 2006. Validating associations in biological databases. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 142–151. ACM.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370.

Grzegorz Kondrak. 2005. N-gram similarity and distance. In *Proceedings of the 12th International Conference String Processing and Information Retrieval*, pages 115–126.

Andre Lamurias, João D Ferreira, and Francisco M Couto. 2015. Improving chemical entity recognition through h-index based semantic similarity. *Journal of Cheminformatics*, 7(Suppl 1):S13.

André Leal, Diogo Gonçalves, Bruno Martins, and Francisco M Couto. 2014. Ulisboa: Identification and Classification of Medical Concepts. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 711–715.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Michael MacCandless, Erik Hatcher, and Otis Gospodnetić. 2010. *Lucene in Action*. Manning Publications Company.

Bridget T McInnes, Ted Pedersen, and Serguei VS Pakhomov. 2009. UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. In *Proceedings of the 2009 Annual Symposium of the American Medical Association*, pages 431–435.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, pages 147–155.

David Sánchez, Montserrat Batet, David Isern, and Aida Valls. 2012. Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications*, 39(9):7718 – 7728.

Joseph Turian, Lev Ratinov, Yoshua Bengio, and Dan Roth. 2009. A preliminary evaluation of word representations for named-entity recognition. In *Proceedings of the NIPS-09 Workshop on Grammar Induction, Representation of Language and Language Learning*, pages 1–8.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A Simple and General Method for Semi-supervised Learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394.

Yaoyun Zhang, Jingqi Wang, Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, and Hua Xu. 2014. Uth_ccb: A Report for Semeval 2014 - Task 7 Analysis of Clinical Text. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 802–806.