ULisboa: Identification and Classification of Medical Concepts

André Leal[+], Diogo Gonçalves[+], 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,dgoncalves}@lasige.di.fc.ul.pt , bruno.g.martins@ist.ul.pt , fcouto@di.fc.ul.pt

Abstract

This paper describes our participation on Task 7 of SemEval 2014, which focused on the recognition and disambiguation of medical concepts. We used an adapted version of the Stanford NER system to train CRF models to recognize textual spans denoting diseases and disorders, within clinical notes. We considered an encoding that accounts with non-continuous entities, together with a rich set of features (i) based on domain specific lexicons like SNOMED CT, or (ii) leveraging Brown clusters inferred from a large collection of clinical texts. Together with this recognition mechanism, we used a heuristic similarity search method, to assign an unambiguous identifier to each concept recognized in the text.

Our best run on Task A (i.e., in the recognition of medical concepts in the text) achieved an F-measure of 0.705 in the strict evaluation mode, and a promising F-measure of 0.862 in the relaxed mode, with a precision of 0.914. For Task B (i.e., the disambiguation of the recognized concepts), we achieved less promising results, with an accuracy of 0.405 in the strict mode, and of 0.615 in the relaxed mode.

1 Introduction

Currently, many off-the-shelf named entity recognition solutions are available, and these can be used to recognize mentions in clinical notes denoting diseases and disorders. We decided to use the Stanford NER tool (Finkel et al., 2005) to train CRF models based on annotated biomedical text.

The use of unsupervised methods for inferring word representations is nowadays also known to increase the accuracy of entity recognition models (Turian et al., 2010). Thus, we also used Brown clusters (Brown et al., 1992; Turian et al., 2009) inferred from a large collection of non-annotated clinical texts, together with domain specific lexicons, to build features for our CRF models.

An important challenge in entity recognition relates to the recognition of overlapping and non-continuous entities (Alex et al., 2007). In this paper, we describe how we modified the Stanford NER system to be able to recognize non-continuous entities, through an adapted version of the SBIEO scheme (Ratinov and Roth, 2009).

Besides the recognition of medical concepts, we also present the strategy used to map each of the recognized concepts into a SNOMED CT identifier (Cornet and de Keizer, 2008). This task is particularly challenging, since there are many ambiguous cases. We describe our general approach to address the aforementioned CUI mapping problem, based on similarity search and on the information content of SNOMED CT concept names.

2 Task and Datasets

Task 7 of SemEval 2014 actually consisted of two smaller tasks: recognition of mentions of medical concepts (Task A) and mapping each medical concept, recognized in clinical notes, to a unique UMLS CUI (Task B). In the first task, recognition of medical concepts, systems have to detect continuous and discontinuous medical concepts that belong to the UMLS semantic group *disorders*. The second task, concerning with normalization and mapping, is limited to UMLS CUIs of SNOMED CT codes (i.e., although the UMLS meta-thesaurus integrates several resources, we are only interested in SNOMED CT). Each concept that was previously recognized can have a unique CUI associated to it, or none at all (CUI-

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 711–715,
Dublin, Ireland, August 23-24, 2014.

LESS). The goal here is to disambiguate the concepts and choose the right CUI for each case. For supporting the recognition and CUI mapping of medical concepts, we retrieved the disorders subset of SNOMED CT directly from UMLS[1].

The evaluation can be done in a strict or a in a relaxed way. For the case of strict evaluation, an exact match must be achieved in the recognition, by having correct start and end offsets, within the text, for the continuous concepts, and a correct set of start and end offsets for the discontinuous concepts. In the relaxed evaluation, there is some space for errors in the offset values from the recognition task. If there is some overlap between the concepts, then the result is considered a partial match, otherwise it is a recognition error.

A set of annotated biomedical texts was given to the participants, separated in three categories: trial, development and training. We also received a final test set, and a large set of non-annotated texts. All the provided texts were initially converted into a common tokenized format, to be used as input to the tools that we considered for developing our approach. After processing, we converted the results back into the format used by SemEval 2014, this way generating the official runs.

3 Entity Recognition

Our entity recognition approach was based on the usage of the Stanford NER software, which employs a linear chain Conditional Random Field (CRF) approach for building probabilistic models based on training data (Finkel et al., 2005). In Stanford NER, model training is based on the L-BFGS algorithm, and model decoding is made through the Viterbi algorithm.

This tool requires all input texts to be tokenized and encoded according to a named entity recognition scheme such as SBIEO (Ratinov and Roth, 2009), characterized by only being able to recognize continuous entities. As we also need to recognize non-continuous entities, we modified the Stanford NER software to use a SBIEON encoding scheme. This new scheme has the following specific token-tag associations:

- **S:** *Single*, that indicates if the token individually constitutes an entity to be recognized.
- **B:** *Begin*, identifying the beginning of the entity. This tag is only given to the first word of the entity, being followed is most cases by tokens labeled as being *inside* the entity.
- **I:** *Inside*, representing the continuation of a non single word entity (i.e., the middle tokens).
- **E:** *Ending*, representing the last word in the case of entities composed by more than one word.
- **N:** *Non-Continuous*, which identifies all the words that are between the beginning and the end of an entity, but that do not belong to it. This label specifically allows us to model the in-between tokens of non-continuous entities.
- **O:** *Other*, which is associated to all other words that are not part of entities.

We developed a Java parser that converts the biomedical text, provided to the participants, into a tokenized format. This tokenized format, in the case of the annotated texts, associates individual tokens to their SBIEON or SBIEO tags, so that the datasets can be used as input to train CRF models.

3.1 Concept Recognition Models

As we said, SBIEON tokenization differs from SBIEO by the fact that the first one gives support to non-continuous entities. Based on these two input schemes, we generated two different models:

Only continuous entities: A 2nd-order CRF model was trained based on the SBIOE entity encoding scheme, which only recognizes continuous entities. Non-continuous and overlapping entities will thus, in this case, only be partially modeled (i.e., we only considered the initial span of text associated to the non-continuous entities).

Non-continuous entities: A 2nd-order CRF model was trained based on the SBIOEN entity encoding scheme, accepting continuous and non-continuous entities, although still not supporting the case of overlapping entities. In these last cases, only the first entity in each of the overlapping groups will be modeled correctly, while the others will only be partially modeled (i.e., by only considering the non-overlapping spans).

Our CRF models relied on a standard set of feature templates that includes (i) word tokens within a window of size 2, (ii) the token shape (e.g., if it is uppercased, capitalized, numeric, etc.), (iii) token prefixes and suffixes, (iv) token position (e.g., at the beginning or ending of a sentence), and (v) conjunctions of the current token with the previous 2 tags. Besides these standard features, we also considered (a) domain-specific lexicons, and (b) word representations based on Brown clusters.

[1] http://www.nlm.nih.gov/research/umls/

3.2 Word Clusters

In addition to the annotated *training* dataset, participants were also provided with 403876 non-annotated texts, containing a total of 828509 tokens. We used this information to induce generalized cluster-based word representations.

Brown et al. proposed a greedy agglomerative hierarchical clustering procedure that groups words to maximize the mutual information of bi-grams (Brown et al., 1992). According to Brown's clustering procedure, clusters are initialized as consisting of a single word each, and are then greedily merged according to a mutual information criterion, based on bi-grams, to form a lower-dimensional representation of a vocabulary that can mitigate the feature sparseness problem. In the context of named entity recognition studies, several authors have previously noted that using these types of cluster-based word representations can indeed result in improvements (Turian et al., 2009). The hierarchical nature of the clustering allows words to be represented at different levels in the hierarchy and, in our case, we considered 1000 different clusters of similar words.

We specifically used the set of training documents, together with the non-annotated documents that were provided by the organizers, to induce word representations based on Brown's clustering procedure, using an open-source implementation that follows the description given by (Turian et al., 2010). The word clusters are latter used as features within the Stanford NER package, by considering that each word can be replaced by the corresponding cluster, this way adding some other *back-off* features to the models (i.e., features that are less sparse, in the sense that they will appear more frequently associated to some of the instances).

4 Disambiguating Concepts

For mapping entities to concept IDs (Task B), we used a heuristic method based on similarity search, supported on Lucene indexes (MacCandless et al., 2010). We look for SNOMED CT concepts that have a high n-gram overlap with the entity name occurring in the text, together with the information content of each SNOMED CT concept.

In our implementation, we used Lucene to retrieve candidate SNOMED CT concepts according to different string distance algorithms: the NGram distance (Kondrak, 2005) first, then according to the Jaro-Winkler distance (Winkler, 1990), and finally according to the Levenshtein distance. The most similar candidate is chosen as the disambiguation. The specific order for the similarity metrics was based on the intuition that metrics based on individual character-level matches are probably not as informative as metrics based on longer sequences of characters, although they can be useful for dealing with spelling variations. However, for future work, we plan to explore more systematic approaches (e.g., based on learning to rank) for combining multiple similarity metrics.

Additionally to the aforementioned similarity metrics, a measure of the Information Content (IC) of each SNOMED CT concept was also employed, to further disambiguate the mappings (i.e., to select the SNOMED CT identifier that is more general, and thus more likely to be associated to a particular concept descriptor). Notice that the IC of a concept corresponds to a measure of its specificity, where higher values correspond to more specific concepts, and lower values to more general ones. Given the frequency $\mathrm{freq}(c)$ for each concept c in a corpus (i.e., the same corpus that was used to infer the word clusters), the information content of this concept can be computed from the ratio between its frequency (including its descendants) and the maximum frequency of all concepts (Resnik, 1995):

$$\mathrm{IC}(c) = -\log\left(\frac{\mathrm{freq}(c)}{\mathrm{maxFreq}}\right)$$

In the formula, $\mathrm{maxFreq}$ represents the maximum frequency of a concept, i.e. the frequency of the *root* concept, when it exists. The frequency of a concept can be computed using an extrinsic approach that counts the exact matches of the concept names on a large text corpus.

5 Evaluation Experiments

We submitted three distinct runs to the SemEval competition. These runs were as follows:

Run 1: A SBIOEN model was used to recognize non-continuous entities. This model was trained using only the annotated texts from the provided training set. We also used some domain specific lexicons like SNOMED CT, or lists with names for drugs and diseases retrieved from DBPedia. Finally, the recognition model also used Brown clusters generated from the non-annotated datasets provided in the competition.

For assigning a SNOMED CT identifier to each entity, we used the disambiguation technique supported by Lucene indexes. In this specific run we used all the considered heuristics for similarity search.

Run 2: A simpler model based on the SBIOE scheme was used in this case, which can only recognize continuous entities. The same features from Run 1 were used for training the recognition model.

For assigning the SNOMED CT identifier to each entity, we also used the same strategy that was presented for Run 1.

Run 3: A similar SBIOE model to that from Run 2 was used for the recognition.

For assigning the corresponding SNOMED CT identifier to each entity, we in this case limited the heuristic rules that were used. Instead of using the string similarity algorithms, we used only exact matches, together with the information content measure and the neighboring terms for disambiguation.

6 Results and Discussion

We present our official results in Table 1, which highlights our best results for each task.

Specifically in Task A, we achieved encouraging results. Run 1 achieved an F-measure of 0.705 in the strict evaluation, and of 0.862 in the relaxed evaluation. Since Runs 2 and 3 used the same recognition strategy (i.e., models that attempted to capture only the continuous entities), we obtained the same results for Task A in both these runs. Table 1 also shows that our performance in Task B was significantly lower than in Task A.

As we can see in the table, our first run was the one with the best results for Task A. The model used on this run recognizes non-continuous entities, and this is perhaps the main reason for the higher results (i.e., the other two runs used the same features for the recognition models).

On what concerns the results of Task B, it is important to notice the distinct results from the first and second runs, which used exactly the same disambiguation strategy. The differences in the results are a consequence from the use of a different recognition model in Task A. We can see that the ability to recognize non-continuous entities leads to the generation of worse mappings, when considering our specific disambiguation strategy. Our last run is the best in terms of the performance over Task B, but the difference is subtle.

7 Conclusions and Future Work

This paper described our participation in Task 7 of the SemEval 2014 competition, which was divided into two subtasks, namely (i) the recognition of continuous and non-continuous medical concepts, and (ii) the mapping of each recognized concept to a SNOMED CT identifier.

For the first task, we used the Stanford NER software (Finkel et al., 2005), modified by us to recognize not only continuous, but also non-continuous entities. This was possible by introducing the SBIEON scheme, derived from the traditional SBIEO encoding. To increase the accuracy and precision of the recognition we have also used domain specific lexicons and Brown clusters inferred from non-annotated documents.

For the second task, we used a heuristic method based on similarity search, for matching concepts in the text against concepts from SNOMED CT, together with a measure of information content to disambiguate the cases of term polysemy in SNOMED CT. We implemented our disambiguation approach through the Lucene software framework (MacCandless et al., 2010).

In the first task (Task A) we achieved some particularly encouraging results, showing that an off-the-shelf NER system can be easily adapted to the recognition of medical concepts in biomedical text. Our specific modifications to the Stanford NER system, in order to support the recognition of non-continuous entity names, indeed increased the precision and recall on Task A. However, our approach for the disambiguation of the recognized concepts (Task B) performed much worse, achieving an accuracy of 0.615 in the case of the relaxed evaluation. Future developments will therefore focus on improving the component that addressed the entity disambiguation subtask.

Specifically on what regards future work, we plan to experiment with the usage of machine learning methods for the disambiguation subtask, instead of relying on a purely heuristic approach. We are interested in experimenting with the usage of Learning to Rank (L2R) methods, similar to those employed on the DNorm system (Leaman et al., 2013), to optimally combine different heuristics such as the ones that were used in our current approach. A L2R model can be used to rank candi-

	Task A						Task B	
	Strict Evaluation			Relaxed Evaluation				
Run	Precision	Recall	F-measure	Precision	Recall	F-measure	Strict Accuracy	Relaxed Accuracy
1	**0.753**	**0.663**	**0.705**	**0.914**	**0.815**	**0.862**	0.402	0.606
2	0.752	0.660	0.703	0.909	0.806	0.855	0.404	0.612
3	0.752	0.660	0.703	0.909	0.806	0.855	**0.405**	**0.615**

Table 1: Our official results for Tasks A and B of the SemEval challenge focusing on clinical text.

date disambiguations (e.g., retrieved through similarity search with basis on Lucene) according to a combination of multiple criteria, and we can then choose the top candidate as the disambiguation.

Additionally, we plan to use ontology-based similarity measures to validate and improve the mappings (Couto and Pinto, 2013). For example, by assuming that all entities in a given span of text are semantically related with each other, we can use ontology relations to filter likely misannotations (Grego and Couto, 2013; Grego et al., 2013).

Acknowledgments

The authors would like to thank Fundação para a Ciência e Tecnologia (FCT) for the financial support of SOMER (PTDC/EIA-EIA/119119/2010), LASIGE (PEst-OE/EEI/UI0408/2014) and INESC-ID (Pest-OE/EEI/LA0021/2013).

We also would like to thank our colleagues Berta Alves, for her support in evaluating development errors, and Luís Atalaya, for the development of some of the data pre-processing scripts.

References

Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Proceedings of the ACL-07 Workshop on Biological, Translational, and Clinical Language Processing*, pages 65–72.

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 and Helena Sofia Pinto. 2013. The next generation of similarity measures that fully explore the semantics in biomedical ontologies. *Journal of Bioinformatics and Computational Biology*, 11(05):1–11.

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.

Tiago Grego and Francisco M Couto. 2013. Enhancement of chemical entity identification in text using semantic similarity validation. *PloS ONE*, 8(5):1–9.

Tiago Grego, Francisco Pinto, and Francisco Couto. 2013. LASIGE: using conditional random fields and chebi ontology. In *Proceedings of the 7th International Workshop on Semantic Evaluation*, pages 660–666.

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

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.

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.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

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.

William E Winkler. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research of the American Statistical Association*, pages 354–359.