BioinformaticsUA: Concept Recognition in Clinical Narratives Using a Modular and Highly Efficient Text Processing Framework

Sérgio Matos
DETI/IEETA
University of Aveiro
3810-193 Aveiro, Portugal
aleixomatos@ua.pt

Tiago Nunes
DETI/IEETA
University of Aveiro
3810-193 Aveiro, Portugal
tiago.nunes@ua.pt

José Luís Oliveira
DETI/IEETA
University of Aveiro
3810-193 Aveiro, Portugal
jlo@ua.pt

Abstract

Clinical texts, such as discharge summaries or test reports, contain a valuable amount of information that, if efficiently and effectively mined, could be used to infer new knowledge, possibly leading to better diagnosis and therapeutics. With this in mind, the SemEval-2014 Analysis of Clinical Text task aimed at assessing and improving current methods for identification and normalization of concepts occurring in clinical narrative. This paper describes our approach in this task, which was based on a fully modular architecture for text mining. We followed a pure dictionary-based approach, after performing error analysis to refine our dictionaries.

We obtained an F-measure of 69.4% in the entity recognition task, achieving the second best precision over all submitted runs (81.3%), with above average recall (60.5%). In the normalization task, we achieved a strict accuracy of 53.1% and a relaxed accuracy of 87.0%.

1 Introduction

Named entity recognition (NER) is an information extraction task where the aim is to identify mentions of specific types of entities in text. This task has been one of the main focus in the biomedical text mining research field, specially when applied to the scientific literature. Such efforts have led to the development of various tools for the recognition of diverse entities, including species names, genes and proteins, chemicals and drugs, anatomical concepts and diseases. These tools use methods based on dictionaries, rules, and machine learning, or a combination of those depending on the specificities and requirements of each concept type (Campos et al., 2013b). After identifying entities occurring in texts, it is also relevant to disambiguate those entities and associate each occurrence to a specific concept, using an univocal identifier from a reference database such as Uniprot[1] for proteins, or OMIM[2] for genetic disorders. This is usually performed by matching the identified entities against a knowledge-base, possibly evaluating the textual context in which the entity occurred to identify the best matching concept.

The SemEval-2014 Analysis of Clinical Text task aimed at the identification and normalization of concepts in clinical narrative. Two subtasks were defined, where Task A was focused on the recognition of entities belonging to the ‘disorders’ semantic group of the Unified Medical Language System (UMLS), and Task B was focused on normalization of these entities to a specific UMLS Concept Unique Identifier (CUI). Specifically, the task definition required that concepts should only be normalized to CUIs that could be mapped to the SNOMED CT[3] terminology.

In this paper, we present a dictionary-based approach for the recognition of these concepts, supported by a modular text analysis and annotation pipeline.

2 Methods

2.1 Data

The task made use of the ShARe corpus (Pradhan et al., 2013), which contains manually annotated clinical notes from the MIMIC II database[4] (Saeed et al., 2011). The corpus contains 298 documents,

[1] http://www.uniprot.org/

[2] http://www.omim.org/

[3] http://www.ihtsdo.org/snomed-ct/

[4] http://mimic.physionet.org/database.html

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 135–139,
Dublin, Ireland, August 23-24, 2014.

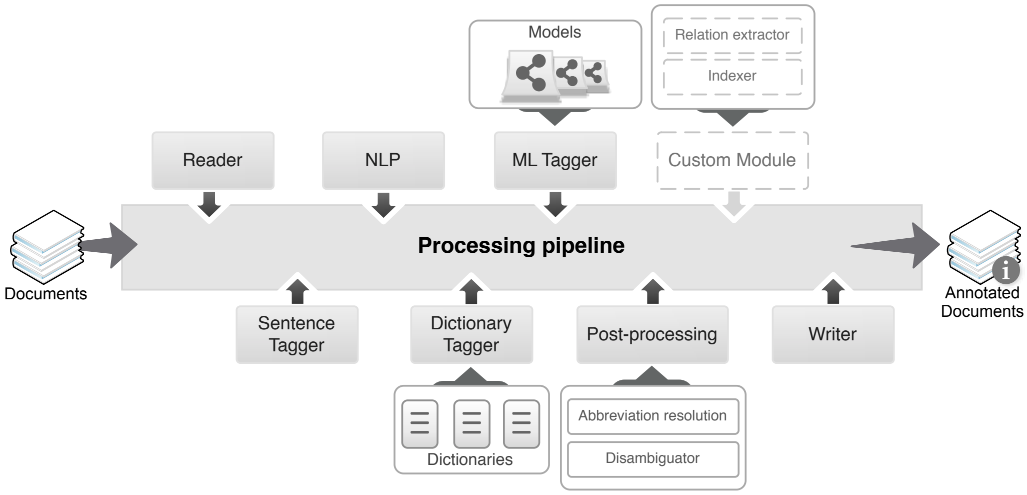

Figure 1: Neji's processing pipeline used for annotating the documents. Boxes with dotted lines indicate optional processing modules. Machine-learning models were not used.

with a total of 11156 annotations of disorder mentions. These annotations include a UMLS concept identifier when such normalization was possible according to the annotation guidelines.

Besides this manually annotated corpus, a larger unannotated data set was also made available to task participants, in order to allow the application of unsupervised methods.

2.2 Processing Pipeline

We used Neji, an open source framework for biomedical concept recognition based on an automated processing pipeline that supports the combined application of machine learning and dictionary-based approaches (Campos et al., 2013a). Apart from offering a flexible framework for developing different text mining systems, Neji includes various built-in methods, from text loading and pre-processing, to natural language parsing and entity tagging, all optimized for processing biomedical text. Namely, it includes a sentence splitting module adapted from the Lingpipe library[5] and a customized version of GDep (Sagae and Tsujii, 2007) for tokenization, part-of-speech tagging, and other natural language processing tasks. Figure 1 shows the complete Neji text processing pipeline, illustrating its module based architecture built on top of a common data structure. The dictionary module performs exact, case-insensitive matching using Deterministic Finite Automatons (DFAs), allowing very efficient processing of documents and matching against dozens of dictionaries containing millions of terms.

Neji has been validated against different biomedical literature corpora, using specifically created machine learning models and dictionaries. Regarding the recognition of disorder concepts, Neji achieved an F-measure of 68% on exact mathing and 83% on approximate matching against the NCBI disease corpus, using a pure dictionary-based approach (Doğan and Lu, 2012).

2.3 Dictionaries

Following the task description and the corpus annotation guidelines, we compiled dictionaries for the following UMLS semantic types, using the 2012AB version of the UMLS Metathesaurus:

- Congenital Abnormality
- Acquired Abnormality
- Injury or Poisoning
- Pathologic Function
- Disease or Syndrome
- Mental or Behavioral Dysfunction
- Cell or Molecular Dysfunction
- Anatomical Abnormality
- Neoplastic Process
- Signs and Symptoms

Additionally, although the semantic type 'Findings' was not considered as part of the 'Disorders' group, we created a customized dictionary including only those concepts of this semantic type that occurred as an annotation in the training data. If

[5] http://alias-i.com/lingpipe/index.html

a synonym of a given concept was present in the training data annotations, we added all the synonyms of that concept to this dictionary. This allowed including some concepts that occur very frequently (e.g. 'fever'), while filtering out many concepts of this semantic type that are not relevant for this task. In total, these dictionaries contain almost 1.5 million terms, of which 525 thousand (36%) were distinct terms, for nearly 293 thousand distinct concept identifiers.

Refining the dictionaries

In order to expand the dictionaries, we preprocessed the UMLS terms to find certain patterns indicating acronyms. For example, if a term such as 'Miocardial infarction (MI)' or 'Miocardial infarction - MI' appeared as a synonym for a given UMLS concept, we checked if the acronym (in this example, 'MI') was also a synonym for that concept, and added it to a separate dictionary if this was not the case. This resulted in the addition of 10430 terms, for which only 1459 (14%) were distinct, for 2086 concepts. These numbers reflect the expected ambiguity in the acronyms, which represents one of the main challenges in the annotation of clinical texts.

Furthermore, in order to improve the baseline results obtained with the initial dictionaries, we performed error analysis to identify frequent errors in the automatic annotations. Using the manual annotations as reference, we counted the number of times a term was correctly annotated in the documents (true positives) and compared it to the number of times that same term caused an annotation to be incorrectly added (a false positive). We then defined an exclusion list containing 817 terms for which the ratio of these two counts was 0.25 or less.

Following the same approach, we created a second exclusion list by comparing the number of FNs to the number of FPs, and selecting those terms for which this ratio was lower than 0.5. This resulted in an exclusion list containing 623 terms.

We also processed the unannotated data set, in order to identify frequently occurring terms that could be removed from the dictionaries to avoid large numbers of false positives. This dataset includes over 92 thousand documents, which were processed in around 23 minutes (an average of 67 documents per second) and produced almost 4 million annotations. Examples of terms from our dictionaries that occur very frequently in this data set are: 'sinus rhythm', which occurred almost 35 thousand times across all documents, and 'past medical history', 'allergies' and 'abnormalities', all occurring more than 15 thousand times. In fact, most of the highly frequent terms belonged to the 'Findings' semantic type. Although this analysis gave some insights regarding the content of the data, its results were not directly used to refine the dictionaries, since the filtering steps described above led to better overall results.

2.4 Concept Normalization

According to the task description, only those UMLS concepts that could be mapped to a SNOMED CT identifier should be considered in the normalization step, while all other entities should be added to the results without a concept identifier. We followed a straightforward normalization strategy, by assigning the corresponding UMLS CUIs to each identified entity, during the dictionary-matching phase. We then filtered out any CUIs that did not have a SNOMED CT mapping in the UMLS data. In the cases when multiple idenfiers were still left, we naively selected the first one, according the dictionary ordering defined above, followed in the end by the filtered 'Findings' dictionary and the additional acronyms dictionary.

3 Results and Discussion

3.1 Evaluation Metrics

The common evaluation metrics were used to evaluate the entity recognition task, namely $Precision = TP/(TP + FP)$ and $Recall = TP/(TP+FN)$, where TP, FP and FN are respectively the number of true positive, false positive, and false negative annotations, and $Fmeasure = 2 \times Precision \times Recall/(Precision + Recall)$, the harmonic mean of precision and recall. Additionally, the performance was evaluated considering both strict and relaxed, or overlap, matching of the gold standard annotations.

For the normalization task, the metric used to evaluate performance was accuracy. Again, two matching methods were considered: strict accuracy was defined as the ratio between the number of correct identifiers assigned to the predicted entities, and the total number of entities manually annotated in the corpus; while relaxed accuracy measured the ratio between the number of correct

	Task A						Task B	
	Strict			Relaxed			Strict	Relaxed
Run	P	R	F	P	R	F	Acc	Acc
Best	0,843	0,786	0,813	0,936	0,866	0,900	0,741	0,873
Average	0,648	0,574	0,599	0,842	0,731	0,770	0,461	0,753
0	**0,813**	**0,605**	**0,694**	**0,929**	**0,693**	**0,794**	0,527	**0,870**
1	0,600	0,621	0,610	0,698	0,723	0,710	**0,531**	0,855
2	0,753	0,538	0,628	0,865	0,621	0,723	0,463	0,861

Table 1: Official results on the test dataset. The best results for each task and matching strategy are identified in bold. The best run from all participating teams as well as the overall average are shown for comparison.

identifiers and the number of entities correctly predicted by the system.

3.2 Test Results

We submitted three runs of annotations for the documents in the test set, as described below:

- Run 0: Resulting annotations were filtered using the first exclusion list (817 terms, TP/FP ratio 0.25 or lower). The extra acronyms dictionary was not used, and matches up to 3 characters long were filtered out, except if they were 3 characters long and appeared as uppercase in the original text.

- Run 1: The extra acronyms dictionary was included. The same exclusion list as in Run 0 was used, but short annotations were not removed.

- Run 2: The extra acronyms dictionary was included. The second exclusion list was used, and short annotations were not removed.

Table 1 shows the official results obtained on the test set for each submitted run.

Overall, the best results were obtained with the more stringent dictionaries and filtering, leading to a precision of 81.3% and and F-measure of 69.4%. This results was achieved without the use of the additional acronyms list, and also by removing short annotations. This filtering does not discard annotations with three characters if they appeared in uppercase in the original text, as this more clearly indicates the use of an acronym. Preliminary evaluation on the training data showed that this choice had a small, but positive contribution to the overall results.

We achieved the second-best precision results with this first run, considering both strict and relaxed matching. Although this level of precision was not associated to a total loss in recall, we were only able to identify 70% of the disorder entities, even when considering relaxed matching. To overcome this limitation, we will evaluate the combined use of dictionaries and machine-learning models, taking advantage of the Neji framework. Another possible limitation has to do with the recognition and disambiguation of acronyms, which we will also evaluate further.

Regarding the normalization results (Task B), we achieved the 12th and 10th best overall results, considering strict and relaxed accuracies respectively, corresponding to the 7th and 6th best team. For relaxed matching, our results are 5,8% lower than the best team, which is a positive result given the naïve approach taken. These performances may be improved as a result of enhancements in the entity recognition step, and by applying a better normalization strategy.

4 Conclusions

We present results for the recognition and normalization of disorder mentions in clinical texts, using a dictionary-based approach . The dictionaries were iteratively filtered following error-analysis, in order to better tailor the dictionaries according to the task annotation guidelines. In the end, a precision of 81.3% was achieved, for a recall of 60.5% and a F-measure of 69.4%. The use of a machine-learning based approach and a better acronym resolution method are being studied with the aim of improving the recall rate.

In the normalization task, using the refined dictionaries directly, we achieved a strict accuracy of 53.1% and a relaxed accuracy of 87.0%. Strict normalization results, as given by the metric defined for this task, are dependent on the entity recognition recall rate, and are expected to follow improvements that may be achieved in that step.

Acknowledgements

This work was supported by National Funds through FCT - Foundation for Science and Technology, in the context of the project PEst-OE/EEI/UI0127/2014. S. Matos is funded by FCT under the FCT Investigator programme.

References

David Campos, Sérgio Matos, and José Luís Oliveira. 2013a. A modular framework for biomedical concept recognition. *BMC Bioinformatics*, 14:281.

David Campos, Sérgio Matos, and José Luís Oliveira, 2013b. *Current Methodologies for Biomedical Named Entity Recognition*, pages 839–868. John Wiley & Sons, Inc., Hoboken, New Jersey.

Rezarta Islamaj Doğan and Zhiyong Lu. 2012. An improved corpus of disease mentions in PubMed citations. In *Proceedings of BioNLP'12*, pages 91–99, Stroudsburg, PA, USA, June.

Sameer Pradhan, Noemie Elhadad, Brett South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy Chapman, and Guergana Savova. 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. *Online Working Notes of the CLEF 2013 Evaluation Labs and Workshop*.

Mohammed Saeed, Mauricio Villarroel, Andrew Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin Kyaw, Benjamin Moody, and Roger Mark. 2011. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. *Critical Care Medicine*, 39(5):952.

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with LR models and parser ensembles. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1044–1050, Prague, Czech Republic.