BioinformaticsUA: Machine Learning and Rule-Based Recognition of Disorders and Clinical Attributes from Patient Notes

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

José Sequeira
DETI/IEETA
University of Aveiro
3810-193 Aveiro, Portugal
sequeira@ua.pt

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

Abstract

Natural language processing and text analysis methods offer the potential of uncovering hidden associations from large amounts of unprocessed texts. The SemEval-2015 Analysis of Clinical Text task aimed at fostering research on the application of these methods in the clinical domain. The proposed task consisted of disorder identification with normalization to SNOMED-CT concepts, and disorder attribute identification, or template filling.

We participated in both sub-tasks, using a combination of machine-learning and rules for recognizing and normalizing disease mentions, and rule-based methods for template filling. We achieved an F-score of 71.2% in the entity recognition and normalization task, and a slot weighted accuracy of 69.5% in the template filling task.

1 Introduction

Biomedical text mining offers the promise of leveraging the huge amounts of information available on scientific documents to help raise new hypotheses and uncover hidden knowledge. Biomedical text mining (TM) has been an important focus of research during the last years, sustained by the high volumes of data, the diverse computational and multi-disciplinary challenges posed, and by the potential impact of new discoveries (Simpson and Demner-Fushman, 2012). These benefits have been demonstrated in recent studies in which text mining methods were used to suggest biomarkers for diagnosis and for measuring disease progression, targets for new drugs, or new uses for existing drugs (Frijters et al., 2010). Likewise, clinical information stored as natural language text in discharge notes and reports could be exploited to identify important associations, and this has led to an increased interest in applying text mining techniques to such texts, in order to extract information related to diseases, medications, and adverse drug events, for example (Zhu et al., 2013).

Research efforts in biomedical text mining have led to the development of various methods and tools for the recognition of diverse entities, including species names, genes and proteins, chemicals and drugs, anatomical concepts and diseases. These methods are based on dictionaries, rules, and machine learning, or a combination of those depending on the specificities and requirements of each concept type. After identifying entity mentions in text, it becomes necessary to perform entity normalization, which consists in assigning a specific concept identifier to each entity. 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.

Following up on the 2014 task, in which the objective was the identification and normalization of disease concepts in clinical texts (Pradhan et al., 2014), two subtasks were defined for the SemEval-2015 Analysis of Clinical Text task. Task 1 consisted of recognizing concepts belonging to the 'disorders' semantic group of the Unified Medical Language System (UMLS) and normalizing to the

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 422–426,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

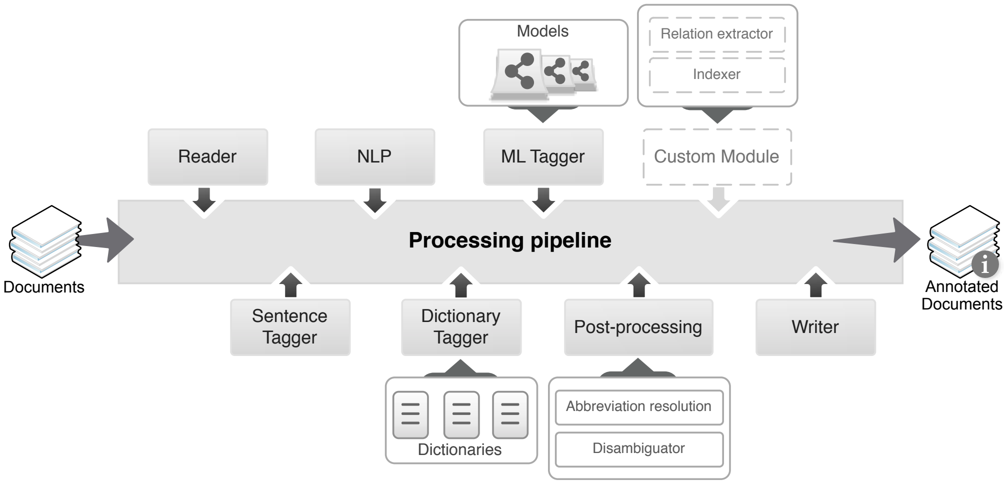

Figure 1: Neji's processing pipeline used for annotating the documents. Dashed boxes indicate optional modules.

SNOMED CT[1] terminology, and Task 2 consisted of identifying and normalizing specific attributes for each disorder mention, including negation, severity, and body location, for example. The task made use of the ShARe corpus (Pradhan et al., 2013), which contains manually annotated clinical notes from the MIMIC II database[2] (Saeed et al., 2011). The task corpus comprised 531 documents, divided into a training portion with 298 documents, a development portion with 133 documents, and a test portion with 100 documents.

In this paper, we present a combined machine-learning and rule-based approach for these tasks, supported by a modular text analysis and annotation pipeline.

2 Methods

Our approach consists of three sequential steps, namely: entity recognition, rule-based span adjustment and normalization, and rule-based template filling. For entity recognition we used Gimli (Campos et al., 2013b), an open-source tool for training machine learning (ML) models that includes simple configuration of the feature extraction process, and Neji, a framework for biomedical concept recognition, integrating modules for natural language processing (NLP) and information extraction (IE), specially tuned for the biomedical domain (Campos et al., 2013a). Figure 1 shows the complete processing pipeline.

2.1 Entity Recognition

We applied a supervised machine-learning approach, based on Conditional Random Fields (CRFs) (Lafferty et al., 2001; McCallum, 2002). The BIO (Beginning, Inside, Outside) scheme was used to encode the entity annotations. To select the best combination of features, we performed backward feature elimination using the supplied training and development data to create and evaluate the models. We then used all the data to train a first-order CRF model with the final feature set, which consisted of the following features:

- NLP features:
  - Token and lemma
- Orthographic features:
  - Capitalization (e.g., "StartCap" and "AllCaps");
  - Digits and capitalized characters counting (e.g., "TwoDigit" and "TwoCap");
  - Symbols (e.g., "Dash", "Dot" and "Comma");
- Morphological features:

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

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

  - Suffixes and char n-grams of 2, 3 and 4 characters;

- Local context:
  - Conjunctions of lemma and POS features, built from the windows {-1, 0}, {-2, -1}, {0, 1}, {-1, 1} and {-3, -1} around the current token.

Apart from the ML model, documents were also annotated with dictionaries for the UMLS 'Disorders' semantic group and a specially compiled acronyms dictionary, as used in the 2014 edition of the task (Matos et al., 2014). In total, these dictionaries contain almost 1.5 million terms, of which 525 thousand (36%) are distinct terms, for nearly 293 thousand distinct concept identifiers. Including this dictionary-matching step produced a small improvement in terms of F-score.

2.2 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. To achieve this step, we indexed the terms of the UMLS concepts that included a SNOMED-CT identifier in a Solr [3] instance. Additionally, we also indexed each term that occurred in the training and development data, together with the corresponding identifier.

To perform normalization of an identified entity mention, we follow a series of steps. First we search the index for the exact term and, if it is found as a gold-standard annotation on the training data, we assign the same identifier to the new mention. If multiple identifiers were used on the training data for the same term, we keep the most commonly assigned one. If the exact mention is not found on the training data, we try to remove a set of 162 prefix (e.g. 'chronic', 'acute', 'large') and 48 suffix terms (e.g. 'changes', 'episodes') obtained from an error analysis on the development data. We then look for this adjusted term on the gold standard annotations and on the UMLS concept synonyms, and use the corresponding identifier and the adjusted mention span. Finally, we try to expand the term to include anatomical regions occurring before or after the identified disorder mention, in order to identify more specific concepts. If such a concept is found on the index, the corrected span is used, together with the corresponding identifier.

2.3 Template Filling

This subtask consists of identifying various attributes of the disorders, such as negation or uncertainty, and normalizing their values according to the nomenclature specified by the task. To address this task, we followed a rule-based approach. For each type of attribute, or slot, we compiled the cue words and the corresponding normalized value from the training and development data. We then created patterns, implemented through regular expressions, to locate these possible cues in the vicinity of each disorder term. To apply the regular expressions, we replace each entity mention in the texts by a generic placeholder, adjusting the cue word spans accordingly when a match is found. For example, to fill the 'Severity' attribute we look for the occurrence of a cue word, associated to this attribute in the training data, that occurs up to *n*[4] characters before or after a disorder mention. This can be expressed by the following regular expression, in which only two alternative cue words are shown for brevity:

```
(mild|sharp|...)\s.{0,15}?__DISO__ |
__DISO__\s.{0,15}?(mild|sharp|...)
```

3 Results and Discussion

3.1 Evaluation Metrics

Task 1 was evaluated by strict and relaxed F-scores. In the first case, the identified text span has to be exactly the same as the gold-standard annotation, and the predicted concept identifier has to match the gold annotation. In the second case, a prediction is considered a true-positive is there is any word overlap between the predicted span and the gold-standard, as long as the identified is correctly predicted.

Task 2 was evaluated in terms of weighted accuracy, which is calculated using a pre-assigned weight for each slot based on its prevalence in the training set.

[3] http://lucene.apache.org/solr/

[4] *n* was empirically set as 5 for the body location attribute, and 15 for all other attributes

	Task 1 performance (P / R / F)			
	Development		Test	
Run	Strict	Relaxed	Strict	Relaxed
1	48.1 / 54.4 / 51.0	51.8 / 58.0 / 54.7	0.669 / 0.738 / 0.702	0.698 / 0.769 / 0.732
2	62.3 / 70.6 / **66.2**	67.5 / 74.7 / **70.9**	0.690 / 0.736 / **0.712**	0.719 / 0.766 / **0.742**
3	62.3 / 70.5 / 66.1	67.4 / 74.5 / 70.8	0.691 / 0.735 / **0.712**	0.720 / 0.765 / **0.742**

Table 1: Development results and official results on the test dataset, for Task 1. P: Precision; R: Recall; F: F-score.

3.2 Test Results

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

- Run 1: In this run, the identified disorder mentions were not first checked against the training data annotations;
- Run 2: The identified disorder mentions were first checked against the training data annotations and the corresponding identifier was used;
- Run 3: Same as Run 2, but the machine learning model was trained only on discharge documents, that is, other document types were not used in the training.

Table 1 shows the results obtained on the development set, and the official results obtained on the test set for each submitted run in Task 1.

As can be observed from the results, using the identifiers assigned in the training data for disease mentions that re-occur in the test data has a very positive impact on the results, increasing precision by 2%. Although this approach may be considered to artificially improve the results, the rationale for using it is that human annotators tend to re-use the same identifier in the case of a ambiguous term. The same might also be true for clinical coders when processing the patient notes.

Comparing our results to the best submitted runs, we verify that we obtain the best recall rates when considering both strict and relaxed scores, but with a significant drop in precision when compared to those results.

Figure 2 illustrates the results obtained on the template filling task. We achieved a slot weighted accuracy of 69.5%. Comparing the results, we achieved the best accuracy for the disease CUI slot. On the other hand, we achieved considerable lower accuracies on the body location and conditional slots, when compared to the top performing runs.

4 Conclusions

We present results for the recognition, normalization and template filling of disorder concepts in clinical texts, using a machine-learning and rule-based approach. We achieved a strict F-score of 71.2% and a relaxed F-score of 74.2%, and obtained the best recall under both evaluation modes. One of the reasons for the lower precision is related to the normalization method. As future work, we will continue developing this step.

We applied a simple rule-based approach for the template filling task, and achieved a weighted accuracy of 69.5%. We aim to continue improving this information extraction step, by acquiring a larger set of possible cue words and revising some of the extraction rules.

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 L. Oliveira. 2013a. A modular framework for biomedical concept recognition. *BMC Bioinformatics*, 14:281.

David Campos, Sérgio Matos, and José L. Oliveira. 2013b. Gimli: open source and high-performance biomedical name recognition. *BMC bioinformatics*, 14:54.

Raoul Frijters, Marianne van Vugt, Ruben Smeets, Ren C. van Schaik, Jacob de Vlieg, and Wynand

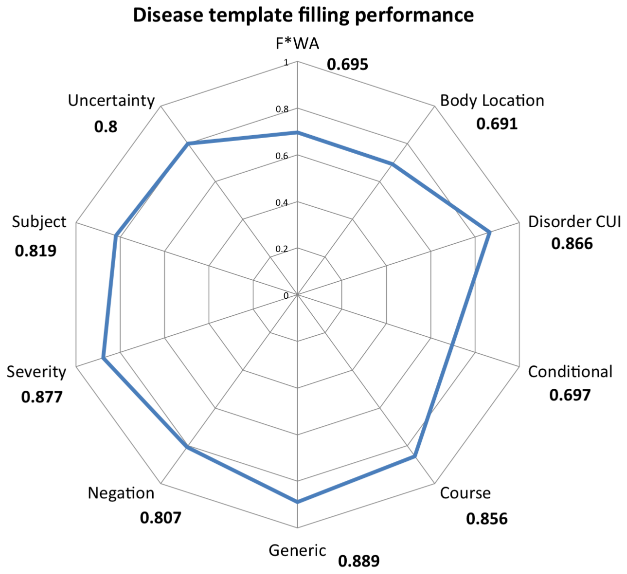

Figure 2: Official results for disease template filling (Task 2).

Alkema. 2010. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Computational Biology*, 6(9):e1000943.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA.

Sérgio Matos, Tiago Nunes, and José L. Oliveira. 2014. BioinformaticsUA: Concept recognition in clinical narratives using a modular and highly efficient text processing framework. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 135–139.

Andrew K. McCallum. 2002. MALLET: A Machine Learning for Language Toolkit.

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*.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. SemEval-2014 Task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland, August.

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.

Matthew S. Simpson and Dina Demner-Fushman. 2012. Biomedical text mining: A survey of recent progress. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 465–517.

Fei Zhu, Preecha Patumcharoenpol, Cheng Zhang, Yang Yang, Jonathan Chan, Asawin Meechai, Wanwipa Vongsangnak, and Bairong Shen. 2013. Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*, 46(2):200–211.