# Extracting relations between diseases, treatments, and tests from clinical data

Oana Frunza and Diana Inkpen

School of Information Technology and Engineering
University of Ottawa, Ottawa, ON, Canada, K1N6N5
{ofrunza,diana}@site.uottawa.ca

**Abstract.** This paper describes research methodologies and experimental settings for the task of relation identification and classification between pairs of medical entities, using clinical data. The models that we use represent a combination of lexical and syntactic features, medical semantic information, terms extracted from a vector-space model created using a random projection algorithm, and additional contextual information extracted at sentence-level. The best results are obtained using an SVM classification algorithm with a combination of the above mentioned features, plus a set of additional features that capture the distributional semantic correlation between the concepts and each relation of interest.

**Keywords:** clinical data-mining, relation classification

## 1 Introduction

Identifying semantic relations between medical entities can help in the development of medical ontologies, in question-answering systems on medical problems, in the creation of clinical trials — based on patient data new trials for already known treatments can be created to test their therapeutic potential on other diseases, and in identifying better treatments for a particular medical case by looking at other cases that followed a similar clinical path. Moreover, identifying relations between medical entities in clinical data can help in stratifying patients by disease susceptibility and response to therapy, reducing the size, duration, and cost of clinical trials, leading to the development of new treatments, diagnostics, and prevention therapies.

While some research has been done on technical data, text extracted from published medical articles, little work has been done on clinical data, mostly because of lack of resources. The data set that we used is the data released in the fourth i2b2-10 shared-task challenges in natural language processing for clinical data[1], the relation identification track in which we participated.

[1] https://www.i2b2.org/NLP/Relations/

## 2 Related Work

The relation classification task represents a major focus for the computational linguistic research community. The domains on which this task was deployed vary wildly, but the major approaches used to identify the semantic relation between two entities are the following: rule-based methods and templates to match linguistic patterns, co-occurrence analysis, and statistical or machine-learning based approaches.

Due to space limitation and the fact that our research is focused on the bioscience domain, we describe relevant previous work done in this domain only using statistical methods.

Machine learning (ML) methods are the ones that are most used in the community. They do not require human effort to build rules. The rules are automatically extracted by the learning algorithm when using statistical approaches to solve various tasks [1], [2]. Other researchers combined the bag-of-words features extracted from sentences, with other sources of information like part-of-speech [3]. [4] used two sources of information: sentences in which the relation appears and the local context of the entities, and showed that simple representation techniques bring good results.

In our previous work presented in [5], we showed that domain-specific knowledge improves the results. Probabilistic models are stable and reliable for tasks performed on short texts in the medical domain. The representation techniques influence the results of the ML algorithms, but more informative representations are the ones that consistently obtain the best results.

In the i2b2-shared task competition [6] the system that performed the best obtained a micro-averaged F-measure value of 73.65%. The mean of the F-measure scores of all the teams that participated in the competition was 59.58%

## 3 Data Set

The data set annotated with existing relations between two concepts in a sentence (if any) focused on 8 possible relations. These relations can exist only between medical problems and treatments, medical problems and tests, and medical problems and other medical problems.

These annotations are made at sentence level. Sentences that contain these concepts, but without any relation between them, were not annotated. The training data set consisted in 349 records, divided by their type and provenance, while the test set consisted of 477 records. Table 1 presents the class distribution for the relation annotations in the training and the test data. Besides the annotated data, a number of 827 unannotated records were also released.

In order to create training data for the Negative class, a class in which a pair of concepts is not annotated with any relation, we considered sentences that had only one pair of concepts in no relation. This choice yielded in a data set of 1,823 sentences. In the test data set a number of 50,336 pair of concepts was not annotated with a relation. These pairs represent the Negative-class test set. In

| Relation | Training | Test |
|---|---|---|
| PIP (medical problem indicates medical problem) | 1239 | 1,989 |
| TeCP (test conducted to investigate medical problem) | 303 | 588 |
| TeRP (test reveals medical problem) | 1734 | 3,033 |
| TrAP (treatment is administered for medical problem) | 1423 | 2,487 |
| TrCP (treatment causes medical problem) | 296 | 444 |
| TrIP (treatment improves medical problem) | 107 | 198 |
| TrNAP (treatment is not administered because of medical problem) | 106 | 191 |
| TrWP (treatment worsens medical problem) | 56 | 143 |

**Table 1.** The number of sentences of each relation in the training and test data sets.

the entire training data a number of 6,381 sentences contained more than two concepts. In the test data this number raised to 10,437.

## 4 Method description

Our method is using a supervised machine learning setting with various types of feature representation techniques.

### 4.1 Data representation

The features that we extracted for representing the pair of entities and the sentence context use lexical information, information about the type of concept of each medical entity, and additional contextual information about the pair of medical concepts.

*The bag-of-words (BOW)* feature representation uses single token features with a frequency-based representation.

*ConceptType* The second type of features represents semantic information about the type of medical concept of each entity: problem, treatment, and test.

*ConText* The third type of feature represents information extracted with the ConText tool [7]. The system is capable to provide three types of contextual information for a medical condition: Negation, Temporality, and Experiencer.

*Verb phrases* In order to identify verb phrases, we used the Genia tagger[2] tool. The verb-phrases identified by the tagger are considered as features. We removed the following punctuation marks: [ . , ’ ( ) # $ % & + * / = < > [ ] - _ ], and considered valid features only the lemma-based forms of the identified verb-phrases.

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

*Concepts* In order to make use of the fact that we know what token or sequence of tokens represents the medical concept, we extracted from all the training data a list of all the annotated concepts and considered this list as possible nominal values for the *Concept* feature.

*Semantic vectors* Semantic vector models are models in which concepts are represented by vectors in some high dimensional space. Similarity between concepts is computed using the analogy of similarity or distance between points in this vector space. The main idea behind semantic vector models is that words and concepts are represented by points in a mathematical space, and this representation is learned from text in such a way that concepts with similar or related meanings are near to one another in that space.

In order to create these semantic vectors and use them in our experiments we used the Semantic Vectors Package[3] [8]. The package uses indexes created by applying a Random Projection algorithm to term-document matrices created using Apache Lucene[4].

We used the semantic vectors to extract the top 300 terms correlated with each relation and to determine the semantic distribution of a pair of concepts in the training corpus of all 9 relations.

## 5 Classification technique

As classification algorithms, we used the SVM implementation with polynomial kernel from the Weka[5] tool.

To solve the task, we are using a 9-class classification model, 8 relations of interest and the Negative class, and also a model that uses a voting ensemble of 8 binary classifiers. The ensemble consists of 8 binary classifier focused on one of the relations and the Negative class. We identify the negative test instances when we use the voting ensemble as being the data points that are classified as Negative by all 8 binary classifiers. Once these negative instances are eliminated, we deploy an 8-class classifier to identify the relations that exist between the remaining instances.

## 6 Results

In this section, we present the results obtained in the competition and post-competition experimental results. The evaluation metric is micro-averaged F-measure.

Table 2 presents our results on the test data, both the competition results and the post-competition ones. More details on the competition experiments can be found in [9].

[3] http://code.google.com/p/semanticvectors/

[4] http://lucene.apache.org/java/docs/index.html

[5] http://www.cs.waikato.ac.nz/ml/weka/

The post-competition experiments were more mostly focused on capturing the semantic correlation between the terms of the pair of concepts and the instances that are contained in each relation. We also tried to capture the verb-phrases overlap between the training and test instances, because these relations evolve around the verbs that are attached to the concept pair. As we can see from Table 2, the post-competition results improved the competition results and the best representation technique is the one that uses a combination of BOW, semantic vectors information, type of the concepts, and verb phrases.

| **Competition** | |
|---|---|
| BOW + Concept + ConceptType + ConText | 40.88% |
| **BOW + ConceptType** | **40.98%** |
| BinaryClassifiers | 39.34% |
| **Post-competition** | |
| SemVect_300 | 40.49% |
| SemVect+VPs+ConceptType | 44.44% |
| BOW + SemVect + VPs + ConceptType | 47.05% |
| BOW + SemVect + VPs + ConceptType + DistSem | 47.53% |
| **BOW(context) + ConceptType + VPs + DistSem + VBs** | **86.15%** |

**Table 2.** F-measure results in the competition.

## 7 Discussion and Conclusions

The results obtained in the competition showed that a richer representation better identifies the existing relations. The ensemble of classifiers showed more balance between all the measures. Since the ensemble of classifiers showed promising results in weeding out the negative examples, we run more experiments when using only 8 relations of interest. With this setting, we obtain the best result of 86.15%. In this experiment, we used additional nominal features for each relation containing verbs that are synonyms to the verbs that describe each relation. The value of these features is the number of verbs overlapping with the context of each pair. The contexts consist in all the words all the words between the pair. The features that we used are presented in Table 2.

We believe that the results can be further improved by using classifiers that are trained on the relations that exist between a certain type of concepts, e.g., one classifier that is trained only on the relations that exist between medical problems and treatments, etc. Our post-competition results are exceeding the mean results in the competition.

As future work, we plan to focus more on adding features that are specific for each concept, reduce the context from sentence level to shorter contexts, look into more verb information, and better understand and incorporate additional information for each relation.

## References

1. Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B.: Prebind and textomy: Mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics **4**(11) (2003) 11–24
2. Mitsumori, T., Murata, M., Fukuda, Y., Doi, K., Doi, H.: Extracting protein-protein interaction information from biomedical text with svm. IEICE Transactions on Information and Systems **89**(8) (2006) 2464–2466
3. Bunescu, R., Mooney, R.: shortest path dependency kernel for relation extraction. In: In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP). (2005) 724–731
4. Giuliano, C., Lavelli, A., Romano, L.: Exploiting shallow linguistic information for relation extraction from biomedical literature. In: 11th Conference of the European Chapter of the Association for Computational Linguistics. (2006) 401–409
5. Frunza, O., Inkpen, D., Tran, T.: A machine learning approach for identifying disease-treatment relations in short texts. IEEE Transactions on Knowledge and Data Engineering **in press** (2010)
6. Roberts, K., Rink, B., Harabagiu, S.: Extraction of medical concepts, assertions, and relations from discharge summaries for the fourth i2b2/va shared task. (2010)
7. Chapman, W., Chu, D., Dowling, J.: Context: an algorithm for identifying contextual features from clinical text. In: In: ACL07 workshop on Biological, translational, and clinical language processing (BioNLP07). (2007) 81–88
8. Widdows, D., Ferraro, K.: Semantic vectors: a scalable open source package and online technology management application. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, B.M.J.M.J.O.S.P.D.T., ed.: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, European Language Resources Association (ELRA) (may 2008) http://www.lrec-conf.org/proceedings/lrec2008/.
9. Frunza, O., Inkpen, D.: Identifying and classifying semantic relations between medical concepts in clinical data (i2b2 challenge). (2010)