# Identifying and classifying semantic relations between medical concepts in clinical data (I2b2 Challenge)

**Oana Frunza, M.Sc., Diana Inkpen, Ph.D.**
**University of Ottawa, ON, Canada**

## Abstract

*In this paper, we describe the three system runs that we submitted to the I2B2-10 Shared Task Challenges in Natural Language Processing and Clinical Data. We participated in the relation identification track of the competition. Our models use a combination of lexical representation, medical semantic information, and additional contextual knowledge in combination with SVM classification algorithms. The best results on the test set are obtained by a 9-class classification algorithm using all types of features as representation technique.*

## Introduction

The I2B2-10 Shared-Task Challenges in Natural Language Processing for Clinical Data is focused on three tasks: extraction of medical problems, tests, and treatments; classification of assertions made on medical problems; and relations between medical problems, tests, and treatments.

The data set released in the competition includes discharge summaries from Partners HealthCare and from Beth Israel Deaconess Medical Center (MIMIC II Database[1]), as well as discharge summaries and progress notes from University of Pittsburgh's Medical Center. All the records have been fully de-identified and manually annotated for concept, assertion, and relation information.

For a period of three months, the training data mentioned above was released to the registered teams for developing their fully-automatic systems. When the test data was released, the teams were required to submit their systems' results for the tracks they registered in.

In the next section, we briefly describe the tasks of the competition, with more emphasis on the relation identification, the task that we registered for.

## Tasks Description

As mentioned earlier, the fourth shared I2B2 challenge had three tracks, all evolving around the medical concepts in clinical data.

### *Concept annotation*

The teams that registered for this track were asked to identify medical concepts that are represented by complete noun phrases (NPs) and adjective phrases (APs), following some guideline constraints.

The concepts of interest are:
*Medical Problems,* represented by phrases that contain observations made by patients or clinicians about the patient's body or mind that are thought to be abnormal or caused by a disease.
*Treatments,* represented by phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem. They are loosely based on the UMLS[2] semantic types therapeutic or preventive procedure, medical device, steroid, pharmacologic substance, biomedical or dental material, antibiotic, clinical drug, and drug delivery device.
*Tests,* represented by phrases that describe procedures, panels, and measures that are done to a patient or a body fluid or sample, in order to discover, rule out, or find more information about a medical problem.

### *Assertion annotation*

This task involves classifying each medical problem into an assertion category. Each medical problem will be assigned to one of six categories of assertions.
The assertion categories are: *present, absent, possible, conditional, hypothetical, and not associated with the patient.*

### *Relation annotation*

The relation track of the competition required the teams to determine the type of relationship that exists between two concepts in a sentence (if any). Relations can exist only between medical problems and treatments, medical problems and tests, and medical problems and other medical problems.

There were a total of 8 possible relation types between these medical concepts:

a. treatment improves medical problem (TrIP);

[1] http://www.physionet.org/pn5/mimic2db/

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

b. Treatment worsens medical problem (TrWP);

c. Treatment causes medical problem (TrCP);

d. Treatment is administered for medical problem (TrAP);

e. Treatment is not administered because of medical problem (TrNAP);

f. Test reveals medical problem (TeRP);

g. Test conducted to investigate medical problem (TeCP);

h. Medical problem indicates medical problem (PIP).

These annotations are made at sentence level. Sentences that contain these concepts, but without any relation between them were not annotated.

In the relation annotation task information coming from concept identification and assertion annotations was available for use.

## Data set

The training data set consisted in 349 records, divided by their type and provenance. Table 1 presents this information. Table 2 presents the class distribution for the relation annotations in the training data.

| Partners | Beth-Israel Deaconess Med Center | University of Pittsburgh Med Center | |
|---|---|---|---|
| Discharge Summaries (DS) | Discharge Summaries (DS) | DS | Patient Notes |
| 97 | 73 | 98 | 81 |

**Table 1.** Summary of the training data set (the columns describe the providing medical institution).

| Relation Type | Sentences |
|---|---|
| PIP | 1003 |
| TeCP | 235 |
| TeRP | 1305 |
| TrAP | 1121 |
| TrCP | 229 |
| TrIP | 90 |
| TrNAP | 85 |
| TrWP | 49 |

**Table 2.** The number of relations of each kind, from the training data set

## Method description

We participated in the competition with three system runs for the relation identification track. Two runs used the same classification approach, with different representation techniques, while the third one used an ensemble of classifiers.

### *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 that are delimitated by spaces. The corpus was pre-tokenized and each token space-separated. No tokens were removed, since we work with short texts and therefore we decided to keep all the tokens. We used a frequency representation in the BOW.

*The second type (ConceptType)* of features represents semantic information about the type of medical concept of each entity. This information is represented for each concept by one of three possible numeric values. These values correspond to each of the possible types of medical concepts: *problem, treatment,* and *test*.

*The third type (ConText)* of feature represents information extracted with the ConText tool, (Chapman et al., 2007). The system is capable to provide three types of contextual information for a medical condition:

(1) **Negation**: ConText determines whether a condition is negated or not.

(2) **Temporality**: ConText can identify if a medical condition is *recent*, *historical* or *hypothetical*.

(3) **Experiencer**: ConText assigns conditions ascribed to someone other than the patient, an Experiencer of *other*.

The tool uses trigger terms, pseudo-trigger terms, and terminations, in order to identify the values for these types of contextual information.

We used six numeric features in order to represent the information provided by ConText: three for each concept in the pair. The first numeric feature flagged if the concept is *affirmed* or *negate;*, the second one had the 3 possible values for temporality; and the third one had either *patient* or *other* as value.

Besides the three types of features described above: *BOW*, semantic-medical information, and *ConText* outputs, we run a few experiments using syntactic information, namely verb-phrases. These results were not among our best results; this is why we decided not to submit any run with them.

In order to identify verb-phrases, we used the Genia tagger[3] tool. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as MEDLINE[4] abstracts.

Figure 1 presents an example of the output of the Genia tagger for the sentence: "*Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of isochamaejasmin.*". The tag O stands for Outside, B for Beginning, and I for Inside.

| | | | | |
|---|---|---|---|---|
| Inhibition | Inhibition | NN | B-NP | O |
| of | of | IN | B-PP | O |
| NF-kappaB | NF-kappaB | NN | B-NP | B-protein |
| activation | activation | NN | I-NP | O |
| reversed | reverse | VBD | B-VP | O |
| the | the | DT | B-NP | O |
| anti-apoptotic | anti-apoptotic | JJ | I-NP | O |
| effect | effect | NN | I-NP | O |
| of | of | IN | B-PP | O |
| isochamaejasmin | isochamaejasmin | NN | B-NP | O |
| . | . | . | O | O |

**Figure 1**. Example of Genia tagger output.

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.

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.

When we used a combination of features, from the types mentioned above, we merged each individual feature space to obtain the final vector space.

### *Classification technique*

As classification algorithms, we used the SVM (Cristianini and Taylor, '00) implementation with polynomial kernel from the Weka (Witten and Frank, '05) tool[5]. In the validation experiments performed on the training data sets using 10-fold cross-validation, we also considered the Complement Naïve Bayes (CNB) classifier (Frank and Bouckaert, '06) as a possible solution, since it is known to perform well on imbalanced data sets. Because the SVM classifier's results were consistently better, we have decided to take into consideration only the SVM classifier.

In order to identify which pairs of concepts from a sentence are in a relation and what is the actual relation, we developed a 9-class classification model and a voting ensemble of binary classifiers.

The **9-class classification model** contains the 8 relations of interest and another one that is represented by sentences that have some concepts mentioned, but in no relation. No information needs to be submitted if a pair of concepts co-exists in a sentence in none of the 8 relations. The track required to submit only pairs of concepts that exist in one of the 8 relations.

In order to weed out pairs of concepts that are in no relation, we first used the 9-class classification approach. The 9th class represents the *Negative* class for which an additional data set of 1,823 sentences was created. This data set consists of sentences from the training data in which a single pair of concepts exists and no relation between them was annotated. To create this data set, we used the concept annotation information from the training data. If a sentence has two concepts annotated but no annotation for them was made during the relation annotation, then we considered it to be an instance for the *Negative* class.

Using this model, we can identify if a particular pair of concepts in a sentence co-exists in one of the 8 relations, or in no relation.

**The second model** is based on an ensemble of binary classifiers. We built 8 such classifiers, each corresponding to one of the 8 relations (the classes being a particular relation or not), in combination with the *Negative* class mentioned above. The final decision of this ensemble is taken as follows:

1. if all classifiers classify a pair of concepts in a sentence with the *Negative* class, then the final class is *Negative*, no relation exists between these concepts;
2. if the ensemble of binary classification algorithms do not agree on the Negative class, then a 8-class classification algorithm

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

[4] http://medline.cos.com/

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

is used. This classifier is trained on the 8 semantic relations of interest.

The submission required sentence-level identification of the 8 semantic relations between three types of medical concepts, already annotated in a test set of 477 records.

Before identifying relations between annotated medical concepts, we have to pair concepts that are mentioned in a sentence and create a context for classification. A pair of concepts along with the sentence they were mentioned in represents a test instance. As an observation, the same sentence can represent the context for more than one instance, there are more than 2 concepts annotated in a sentence. A test set of 54,827 instances was created.

### Results on the test data

In this section, we present the results obtained with the two systems on the released test set in the relation identification track of the forth I2B2 Challenge.

The main evaluation metrics used are Micro-averaged Precision, Recall, and F-measure, averaged over all the relation types.

The first run that we submitted represented the output results of the 9-class classification algorithm having: *BOW, Concept* (nominal), *ConceptType*, and *ConText* as features with SVM as classifier.

The second run that we submitted consisted in the deploying the 9-class classification model with *BOW* plus *Concept Type* features and SVM as classifier.

In the third run we used the ensemble of binary classifiers and an 8-class classifier, all based on SVM. The binary classifiers uses *BOW*, *Concept*, and *ConceptType* as features, while the 8-class classifier uses *BOW, ConceptType, and ConText* as features.

Table 3 presents the results for the three runs for all three major evaluation measures for all the relations together.

| Run | Recall | Precision | F-measure |
|---|---|---|---|
| 1 | **61.13%** | 30.71% | 40.88% |
| 2 | 54.64% | **32.79%** | **40.98%** |
| 3 | 59.73% | 29.33% | 39.34% |

**Table 3**. Results on the test set.

### Discussion

The results that we obtained on the test set followed the trend that we observed on the 10-fold cross validation experiments that we did on the training data. The first model that uses all the representation techniques together, has a value for F-measure and precision that is close to the values of the model that uses only the *BOW* and *Concept Type* features. If we look at the recall levels, we can see that the first run is significantly superior to the second one, suggesting that a richer representation better identifies the existing relations.

The third run, that uses the ensemble of classifiers, was close in performance to our best results, showing more balance between all the measures.

We believe that better results could be obtained 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. By deploying a classifier that distinguishes fewer classes and it is focused only on a certain type of relations could increase the chances of identifying the right class and introduce fewer false positive examples.

The classifiers that we submitted were trained on all relations and some of the concept pairs could have been assigned a relation that was not existent. In our models, the *Concept Type* features captured the type of concepts of the pair. An initial step that would triage the pairs of concepts based on the type of medical entities it contains would be a better choice than a classification feature.

### Conclusions and future work

The best results that we obtained on the test set uses an SVM classifier and a rich feature representation space in a 9-class classification task.

As future work, we would like to try experiments where we take into account the observations that we made in the Discussion section.

### References

1. Wendy W. Chapman, David Chu, and John N. Dowling. 2007. *ConText: An Algorithm for Identifying Contextual Features from Clinical Text.* In Proceedings of the ACL Workshop on BioNLP 2007, pages 81–88.
2. Ian H. Witten, and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques* (Second Edition), Morgan Kaufmann, 2005.
3. N. Cristianini and J. Shawe-Taylor. 2000. *An introduction to support vector machines.*, Cambridge University Press, Cambridge, UK 2000.