Information Classification in Clinical NLP – I2b2 challenge

Dung Hoang Minh Nguyen – Supervisor: Professor Jon Patrick
School of IT

FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

1. Introduction

Information Extraction and Classification of clinical data are current challenges in natural language processing with an increased demand for enhancing the accuracy of information extraction. We present a cascaded method to deal with three different extractions from clinical data:

1.Extraction of medical problems, tests, and treatments.

2.Classification of assertions made on medical problems.

3.Relations of medical problems, tests, and treatments.

The outputs of this system are used for evaluation in all three tiers of the Fourth i2b2/VA Shared-Task and Workshop Challenges in Clinical Natural Language Processing 2010.

2. System Architecture

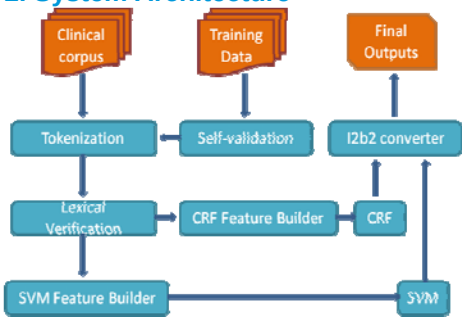

Figure 1. System architecture

Tokenization

Each line (sentence) in the clinical record is split into tokens using white-space tokenization.

Lexical Verification

Lexical verification for each token includes expansion of abbreviations, acronyms, checking against gazetteers and the lexical resources of UMLS, MOBY, SnomedCT, then resolving misspellings and unknown words. All the results of this process are saved in a Lexicon Management System (LMS) for later use in feature generation. The LMS is a system developed to store the accumulated lexical knowledge of our Laboratory and contains categorizations of spelling errors, abbreviations, acronyms and a variety of non-tokens. It also has an interface that supports rapid manual correction of unknown words with a high accuracy clinical spelling suggestor plus the addition of grammatical information and the categorization of such words into gazetteers.

Self-validation of the training data

We identify this as a process of reflexive-validation or alternatively 100% Train and Test. The purpose of this validation is to find the errors in the gold standard itself and the weakest points in our computational methods. Errors in the gold standard were corrected manually so that the model would not learn from incorrect examples. Errors generated by the computational methods were analyzed and the methods improved where possible.

3. Concept annotation

For extraction of medical problems, tests, and treatments, seven feature sets were used for each unigram in the CRF training:

1. Bag of words with context window size of three words.
2. Lemma, part of speech, chunk from the GENIA tagger.
3. Gazetteers and lexical resources (UMLS, MOBY, SCT).
4. Abbreviation, acronym, misspelling expansions.
5. Number of tags (for each token has been recognized as a number, we add a feature with value "number" to distinguish from plain word token).
6. Text to SNOMED Convertor (TTSCT)
7. Medication extraction system's results.

Entity Type	Training	Testing	Recall(test) Recall(train) (Baseline)	Precision(test) Precision (train) (Baseline)	F-score(test) F-score(train) (Baseline)
PROBLEM	11983	18550	79.93% 81.23% (72.28%)	83.53% 84.84% (82.89%)	81.69% 83.00% (77.23%)
TEST	7380	12899	78.94% 80.58% (72.17%)	86.15% 88.14% (88.39%)	82.39% 84.19% (79.46%)
TREATMENT	8515	13560	77.52% 79.05% (65.39%)	85.62% 87.11% (86.60%)	81.37% 82.88% (74.52%)
OVERALL	27878	45009	78.92% 80.39% (70.16%)	84.88% 86.38% (85.38%)	81.79% 83.28% (77.02%)

Table 1. Final scores for concept annotation for the Challenge Test set, 10-fold CV of the training set and baseline.

4. Assertion Classification

For the ASSERTION classification task, the organizers provided the ground-truth for concept annotation. Consequently, the boundary of PROBLEM could be used as a feature for the CRF re-classifier. Six lexicons were manually built corresponding to six properties of PROBLEM (PRESENT, ABSENT, POSSIBLE, CONDITIONAL, HYPOTHETICAL, NOT ASSOCIATED). Each lexicon contained some words or phrases which could contribute to the classification of the assertion. Only four feature types were created for the ASSERTION CRF:

1. Bag of words with a context window size of three words.
2. Lexicon source.
3. Negation identifier.
4. PROBLEM boundary.

The assertion of each problem was based on information in the sentence it belonged to. Generally, the assertion is decided by the nearest word in a lexicon before a PROBLEM, where the name of the lexicon becomes the type of ASSERTION. Priorities of assertion types were also considered, where a new ASSERTION type is assigned only if it has a higher ranking than the current type. If there was no word in any lexicon, the default ASSERTION type was assigned as PRESENT, and there was no

Method	Recall	Precision	F score
RULE-BASED	90.73%	90.73%	90.73%
CRF	92.25%	92.49%	92.37%
SVM	81.77%	81.77%	81.77%

Table 2. The highest scores for Rule-Based, CRF and SVM methods for extracting Assertions from the training set.

Three different methods (rule-based, CRF, SVM) were designed and tested, all of them based on the same ideas of using a lexicon as a key feature to classify the assertions made about medical problems. The comparison was made on the best 10-fold cross-validation results of rule-based, CRF and SVM approaches. Best performance was obtained by using CRF methods.

Assertion Type	Training	Testing	Recall(test) Recall(train)	Precision(test) Precision (train)	F-score(test) F-score(train)
ABSENT	2535	3609	97.19% 94.32%	93.59% 92.93%	93.88% 93.62%
NOT ASSOCIATED	92	145	46.71% 49.65%	78.82% 80.77%	58.76% 58.35%
CONDITIONAL	103	171	18.13% 13.50%	67.39% 70.00%	28.57% 22.76%
HYPOTHETICAL	651	717	69.87% 80.95%	85.06% 91.33%	76.73% 85.83%
POSSIBLE	535	883	49.49% 54.77%	77.48% 79.19%	60.40% 64.75%
PRESENT	8051	13025	97.38% 96.53%	92.51% 93.16%	94.88% 94.82%
OVERALL	11967	18550	91.90% 92.25%	91.90% 92.49%	91.90% 92.37%

Table 3. Scores for Challenge test data and the training set for Assertion classification.

5. Relation Classification

There were nine features used in the SVM to classify the relationships between medical concepts:

1. Three words before the first concept.
2. Three words after the second concept.
3. Words between the two concepts.
4. Words inside of each concept.
5. The type of each concept from the ground-truth.
6. The Assertion type of the PROBLEM concept.
7. Concept types between two concepts.
8. Medication extraction result.
9. Lexicon.

Entity Type	Training	Testing	Recall(test) Recall(train) (Baseline)	Precision(test) Precision (train) (Baseline)	F-score(test) F-score(train) (Baseline)
PIP	1239	1986	62.74% 64.32% (67.95%)	67.68% 72.95% (69.09%)	65.12% 67.91% (65.88%)
InWP	56	143	7.00% 3.57% (3.57%)	50.00% 100% (100%)	9.41% 6.90% (6.90%)
InAP	1422	2487	77.42% 77.92% (77.57%)	69.90% 68.48% (63.68%)	73.35% 72.89% (69.94%)
InNAP	108	161	18.09% 26.47% (25.47%)	55.56% 70.00% (71.05%)	31.19% 38.36% (37.50%)
ToCP	396	444	47.97% 44.59% (47.66%)	49.53% 63.64% (67.95%)	48.74% 52.57% (54.71%)
ToP	107	198	15.66% 23.36% (25.23%)	86.11% 69.44% (64.29%)	26.50% 34.97% (36.24%)
ToCP	303	589	43.03% 47.85% (44.88%)	61.41% 77.13% (74.32%)	50.60% 59.06% (55.97%)
ToRP	1733	3011	84.04% 86.96% (87.31%)	84.04% 82.39% (79.93%)	84.04% 84.62% (83.45%)
OVERALL	5262	9070	67.51% 70.90% (69.8%)	73.07% 74.44% (71.12%)	70.18% 72.63% (70.99%)

Table 4. Scores for Challenge test, training set and a baseline model for relation classification.

6. Future work

We introduced a general NLP system architecture which is easily adapted to different requirements in Clinical Information Extraction and Classification by choosing relevant feature sets.

In future work, more feature sets could be added such as a sentence parse tree. Finally, this system's pipeline will be developed into an Experiment Management System so that researchers could efficiently select various feature sets from a feature list and run the experiment for multiple NLP tasks.

THIS RESEARCH IS SPONSORED BY
Health Information Technologies Research Laboratory (HITRL)

Computing for language, living and life