I2b2 Challenges in Clinical Natural Language Processing 2010

Jon D. Patrick, Dung H.M. Nguyen, Yefeng Wang, Min Li, School of IT, The University of Sydney, NSW 2006, Australia

Abstract

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. In this paper, we present a cascaded method to deal with three different extractions and classifications in clinical data: concept annotation, assertion classification and relation classification. The outputs of this system are used for evaluation in all three tiers of the Fourth i2b2/VA Shared-Task and Workshop Challenges. Overall Concept classification attained an F-score of 83.3% against a baseline of 77.0%, the optimal F-score for Assertions about the concepts was 92.4% using a Conditional Random Field (CRF) machine learner, and an Support Vector Machine (SVM) classifier attained 72.6% for relationships between clinical concepts against a baseline of 71.0%. Micro-average results for the Challenge test set were 81.79%, 91.90%, and 70.18% respectively.

1. Introduction

In the clinical domain, there is a large amount of textual data in patients' notes so efficient processing techniques are necessary to make use of this valuable information. Information Extraction and Classification tools for processing clinical narratives are valuable for assisting clinical staff to quickly finding data of relevance to them.

In this study the focus is on: extraction of medical problems, tests and treatments, classification of assertions made on medical problems, and relationships between medical problems, tests, and treatments[1]. This paper presents an approach for building the required extraction models using both training data and local knowledge resources, including gazetteers of entities, acronyms and abbreviations, and a spelling correction process with methods for resolving unknown words and non-word tokens.

2. Related Work

There are three tasks corresponding to three tiers in the Challenge. Firstly, the concept annotation task (PROBLEM, TEST, TREATMENT) requires Named-Entity Recognition (NER), secondly assertion classification about a Problem with 6 classes (PRESENT, ABSENT, POSSIBLE, CONDITIONAL, HYPOTHETICAL, NOT ASSOCIATED) as a re-classification process, and thirdly the extraction of relations between concepts as a relation classification (RC) task. There are also relationships between the tasks where the output of one task is an input for the next task.

The first combined classifier approach in bioNER proposed a two state-model where boundary recognition and term classification are separated into two phases[2]. In each classification phase, different feature sets were selected independently, which is more efficient for each task.

The closest research related to our methods is the Information Extraction system for clinical notes of Wang[3], where a clinical corpus was annotated for clinical named entities and relationships based on the Systematic Nomenclature of Medicine Clinical Terminology (SNOMED CT), which is a very large logical ontology of clinical terms. This work used CRF and SVM machine learners to create an IE system. Using multiple classifiers rather than a single classifier is the new direction of Wang's NER.

For the RC task, there are many definitions of relationships between concepts where each system classifies different relationship types. In general, relevant features are extracted from the text and are usually selected on the basis of the experimental results and intuition, or by statistical information techniques[4].

In comparison to these IE systems, our work is an adaptation based on the specific requirements of the i2b2 challenge with three levels of classification. We have developed a pipeline system for clinical NLP, which includes a proofreading process, with gold-standard reflexive-validation and correction. The information extraction system is the combination of a machine learning approach and a rule-based approach. Furthermore, a post processing step was implemented to refine the results. Conditional random fields (CRF) and support vector machines (SVM) are two classical machine-learning approaches used, but in this case with some new feature sets (dictionary; abbreviation, acronym, misspelling expansion;

TTSCT[5] and Medication Extraction system's result[6]) introduced.

As the results of other submissions to the Challenge are not available this report only compares the best results of 10-fold cross validation on the training data, and the final results on the Challenge test data as provided by the organisers.

3. Methodology

There are four main steps in our methodology:

1. Proofreading the corpus and training data.
2. Annotation validation.
3. Model building using a customised Language Technology architecture.
4. Use the models to annotate test data and evaluate the final results (see Section 5).

3.1 Proofreading

Proofreading is an important process to clean up the corpus (resolve unknown words, misspelt words) and provide lexical verification for further steps of NLP.

3.2 Annotation Validation

The gold standard needs to be corrected for inconsistencies between concept annotations by reflexive-validation, which we also denote as "100% train & test". This involves using 100% of the training test to build a model and then testing on the same set. As theoretically all data items used for training should be correctly identifiable by the model any errors represent either inconsistencies in annotations or weaknesses in the computational linguistic processing. The former faults identify training items that we reject, and the latter gives us indications on where to concentrate our efforts to improve our processing system.

3.3 Model Building

A 10-fold cross-validation method was used for model building with data from the gold-standard which contained 349 annotated records. After running different experiments with a variety of features and linguistic processes, we generated the model from the training data with the optimal feature set.

4. Language Technology Architecture

The approach for completing the task set was:

1. Use CRF to identify PROBLEMS, TESTS and TREATMENTS.
2. Build and use a dictionary as a feature for the CRF to re-classify PROBLEMS into ASSERTION categories.
3. Build pairs for each concept relationship where at least one concept is the PROBLEM type and classify eight relationship types using SVM.
4. Convert the results of CRF and SVM to i2b2 format.

Figure 1 demonstrates the detailed system architecture, which includes the following processing stages:

I. Tokenization

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

II. 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 categorisations 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 gazeteers[7].

III. Self-validation of the training data

The complete training data set was used to build a model which was then used to annotate the training data to check for differences. This process of reflexive-validation improved scores of the order of .5%.

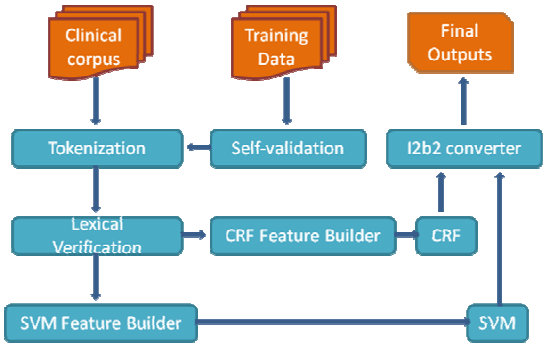

Figure 1. Language Technology Architecture.

IV. CRF feature builder

For CONCEPT annotation, seven feature sets were used for each unigram in the CRF training:

1. Bag of words with a 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 tag. Each token has been recognized as a number, we add a feature with value "number" to distinguish it from word tokens.
6. Text to SNOMED Convertor (TTSCT)[2]
7. Medication extraction system's results[6].

For the ASSERTION classification task, the organisers had provided the ground-truth for concept annotation. Consequently, the boundary of PROBLEM could be used as a feature for the CRF re-classifier. Six lexica were manually built corresponding to the six ASSERTIONs about 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 features types were created for the ASSERTION CRF:

1. Bag of words with a context window size of three words.
2. Lexicon type.
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 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 lexicon compiled for the PRESENT assertion.

V. CRF model training and classifying

The CRF feature builder generated features to train the model. After the learning process, the model was used to classify concepts in the clinical records.

VI. SVM feature builder

For ASSERTION classification task, similar feature sets to CRF methods were generated for SVM classifier. These are:

1. Bag of words with a context window size of five words before and after PROBLEM.
2. Lexicon resource.
3. Negation identifier.
4. Words inside of each PROBLEM.

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 type.

VII. SVM model training and classification

The features which were generated in the previous step were passed to an SVM to build the model. After the learning process, the model was used to classify relationships between concept pairs in clinical records.

VIII. i2b2 output converter

The i2b2 converter converts the output from the CRF and SVM classifiers to the i2b2 format. This produces the final results used for evaluation.

5. Results and Discussion

In this section, the experiment results (baselines, and 10-fold cross-validation results) and final test output for concept annotation, assertion classification and relation classification are presented and discussed.

5.1 Concept Annotation Experiment

This section discusses the extraction of medical PROBLEMS, TESTS and TREATMENTS from the clinical notes. Table 1 illustrates the performance for exact matching using the optimal seven feature sets. The number in brackets is the baseline which used a bag of words as the only feature set. The baseline shows that the most difficult entity to recognize is TREATMENT (F-score of 2.71% lower than PROBLEM and 4.94% lower than TEST).

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.

Other feature sets were used during the experiment process but they did not improve the F-score. These feature sets are morphology and finite-state-automata (FSA) annotation. The best performance was obtained from seven feature sets as described in the CRF Feature Builder section. Each feature set was sequentially added to the CRF builder to train the model and predict the result. If the performance was improved with one feature, this feature was retained, otherwise, it was removed from the training features.

Overall, the best F-score is over 83%, approximately 6% higher than the bag-of-words baseline. TREATMENT still has the lowest score but the difference to PROBLEM (0.12%) and TEST (1.31%) is less significant. This occurs because:

1. There are many ways that TREATMENT can be represented in clinical notes (drug name; drug name with dose; drug name with details in brackets, multiple drug names separated by hyphens, etc.).
2. Misspelling of drug names.
3. Many unseen drug names.

In contrast, the performance for TEST annotation is highest although it has the smallest frequency of the three entity types. The reason is there are fewer varieties of TEST expressions so that the model could learn more effectively.

The 10-fold cross-validation results are nearly 1.5% better than the Challenge test result of 81.79%. This is a loss of due to unseen data, where the total number of concepts in the test set is more than one and a half times greater than training data.

5.2 ASSERTION Classification Experiment

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. In this section, the comparison is made on the best 10-fold cross-validation results of rule-based, CRF and SVM approaches as shown in table 2, and then the best scores amongst them are compared to the Challenge test results in table 3.

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.

The first step was to implement the rule-based method. The F-score performance of the rule-based method is over 90% with a very small lexicon for each class (ABSENT: 27 words, POSSIBLE: 26 words, CONDITIONAL: 7 words, HYPOTHETICAL: 13 words, NOT ASSOCIATED: 8 words, TOTAL: 81 words). With the results of the rule-based method, we considered that a statistical approach based on the same idea would produce a better performance. Consequently, the rule-based method was converted to a statistical method for both CRF and SVMtests. However, only the CRF generated a higher score than the rule-based method (1.64%) while the score of the SVM dropped significantly (nearly 9%). The explanation for this result is:

1. The sequence of words in the sentence and especially before each concept is important in deciding the assertion made about the medical problems.
2. In the CRF method, the sequence of tokens and their features is a key factor to training the model.
3. While for the SVM, only the word itself could be used as feature and so the sequence contributed little to the classification result.

As can be seen from table 2, the best performance was obtained by using CRF methods, while the 2 other methods have the same Recall and Precision scores. The difference in the Precision and Recall of the CRF results are due to minor errors in boundary annotations (with boundary features the CRF model probably combines two consecutive PROBLEMS into one single annotation). This problem could be resolved by a post-processing step which checks for concatenation of PROBLEMS in the CRF output so as to assign Assertion types for each PROBLEM.

Assertion Type	Training	Testing	Recall(test) Recall(train)	Precision(test) Precision (train)	F-score(test) F-score(train)
ABSENT	2535	3609	92.19% 94.32%	93.59% 92.93%	92.88% 93.62%
NOT ASSOCIATED	92	145	46.21% 45.65%	78.82% 80.77%	58.26% 58.33%
CONDITIONAL	103	171	18.13% 13.59%	67.39% 70.00%	28.57% 22.76%
HYPOTHETICAL	651	717	69.87% 80.95%	85.06% 91.33%	76.72% 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.

Table 3 shows the most popular classes (PRESENT, ABSENT) have the highest performance with the F-score greater than 91% in both training and Challenge test data sets. The lowest F-scores were in the scarce types (CONDITIONAL, NOT ASSOCIATED) due to a lack of training examples and small number of words used in their dictionaries, especially CONDITIONAL which performed the worst with just under 30% of F-score.

5.3 Relation Classification Experiment

The support vector machine was used to classify eight relation classes between each PROBLEM and other entities in the sentences, namly, PROBLEM and PROBLEM (PIP); PROBLEM and TEST (TeRP, TECP), PROBLEM and TREATMENT (TrIP, TrWP, TrCP, TrAP, TrNAP). In the specification of the relation classification task, there is no need to indicate if two concepts do not have a relationship. However in the SVM model, no relation was also treated as a type of relationship to enable the classification process. The baselines, best training results and Challenge test results are presented in table 4.

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% (62.95%)	67.68% 72.95% (69.09%)	65.12% 67.91% (65.88%)
TrWP	56	143	2.80% 3.57% (3.57%)	80.00% 100% (100%)	5.41% 6.90% (6.90%)
TrAP	1422	2487	72.46% 77.92% (77.57%)	69.90% 68.48% (63.68%)	71.15% 72.89% (69.94%)
TrNAP	106	191	13.09% 26.42% (25.47%)	55.56% 70.00% (71.05%)	21.19% 38.36% (37.50%)
TrCP	296	444	47.97% 44.93% (47.64%)	49.53% 63.64% (62.95%)	48.74% 52.67% (54.23%)
TrIP	107	198	15.66% 23.36% (25.23%)	86.11% 69.44% (64.29%)	26.50% 34.97% (36.24%)
TeCP	303	588	43.03% 47.85% (44.88%)	61.41% 77.13% (74.37%)	50.60% 59.06% (55.97%)
TeRP	1733	3033	84.04% 86.96% (87.31%)	84.04% 82.39% (79.93%)	84.04% 84.67% (83.45%)
OVERALL	5262	9070	67.51% 70.90% (70.87%)	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.

The baseline result is produced from an SVM classifier with basic feature sets: three words before and after concepts, words between concepts, words inside of each concept, and types of concepts. The

higher the frequency of the relation type the better performance it achieved.

After adding more features (see SVM Feature Builder section) the best result for 10-fold cross-validation increased by 1.64%. The three most frequent classes have the highest F-scores: TeRP (84.62%), TrAP (72.89%) and PIP (67.91%); while the smallest type of TrWP had very low F-scores at under 7%.

The Challenge test data is nearly double the size for the larger classes of the training data set. This causes approximately a 2.5% drop in F-score due to unseen examples.

6. Conclusion

In this paper, a completed system for the i2b2 Clinical NLP challenge has been presented. The system generates results for all three tasks in the Challenge. Furthermore, we also 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 can efficiently select various feature sets from a feature list and run the experiment for multiple NLP tasks.

Acknowledgments

We would like to give a special thanks to members in the Health Information Technologies Research Laboratory for their valuable contributions.

References

1. https://www.i2b2.org/NLP [accessed 09.08.10]

2. Lee, K.J. Hwang, Y.S. and Rim, H.C. Two-phase biomedical NE recognition based on SVMs. In proceedings of the ACL 2003 workshop on Natural language processing in biomedicine 2003; 33-40.

3. Wang, Y. Information Extraction from Clinical Notes. PhD Thesis. School of Information Technologies, University of Sydney 2010.

4. Haddow, B. Using automated feature optimisation to create an adaptable relation extraction system. BioNLP: Current Trends in Biomedical Natural Language Processing 2008; 19-27.

5.Patrick,J. Wang, Y. and Budd, P. An automated system for conversion of clinical notes into SNOMED clinical terminology. Proc. 5rd Australasian symposium on ACSW frontiers 2007; 68:219-226.

6. Patrick, J. and Li, M. A Cascade Approach to Extracting Medication Events. Third i2b2 Shared-Task and Workshop Challenges in Natural Language Processing for Clinical Data, San Francisco, 2009.

7. Patrick, J. Sabbagh, M. Jain, S, and Zheng H. Spelling Correction in Clinical Notes with Emphasis on First Suggestion Accuracy. Proc of 2nd Workshop on building and Evaluating Resources for Biomedical Text Mining (BioTxtM2010). Malta, May 2010.