Concept Identification and Assertion Classification in Patient Health Records

Harsha Gurulingappa[1,2], Martin Hofmann-Apitius[1,2] and Juliane Fluck[1]
[1]Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
Schloss Birlinghoven, 53754, Sankt Augustin, Germany
[2]Bonn-Aachen International Center for Information Technology (B-IT)
Dahlmannstraβe 2, 53113, Bonn, Germany

Abstract

As part of the fourth i2b2/VA challenge 2010, we explored various feature sets and different machine learning based classifiers for automatic medical concept identification and assertion classification in patient health records. For the identification of medical concepts, a Conditional Random Fields based system coupled with a dictionary based named entity recognition system (ProMiner) was applied that achieved F-score of 0.82. For the medical assertion classification, a range of classifiers were evaluated. The highest performance with F-score of 0.9 was achieved with a Support Vector Machine based method.

1. Introduction

The electronic patient health records encompass information about medical problems and therapeutic interventions associated with the patients. Hence, an automatic processing of these health records helps in understanding the underlying causative factors as well as to develop a preventive rationale.

In order to experience the challenges associated with processing the health records, the Fraunhofer SCAI team participated in concept identification and assertion classification challenges put forth by the I2B2 competition (2010). The concept identification task involves the recognition of mentions of medical *problems*, *treatments* and *tests* in the patient health records. The assertion classification task requires an automatic classification of the medical problems into six pre-defined categories such as *present, absent, possible, conditional, hypothetical* or *not associated with the patient*. Definitions and examples of different classes of concepts and assertion types are documented online[1].

There has been a noticeable amount of work related to the identification of biological and medical concepts (*also referred to as named entities*) in free text. They include dictionary-based, rule-based and machine learning-based approaches. For example, ProMiner[2] is a dictionary-based named entity recognition system that has found wide applications in the recognition of biological, medical as well as chemical names in free text. Li et al.[3] proposed a Conditional Random Fields (CRF)-based system for the identification of medical disorders in clinical text. Typical examples of publicly available tools that can be used for the identification of biomedical concepts include the MetaMap program[4] that maps any arbitrary text to the concept in UMLS metathesaurus. MedLEE[5] is another application that can be used for the extraction and encoding of clinical information from patient health narratives.

Once the key concepts in text have been correctly recognized, further processing them such as identifying the assertions or relationships helps in developing the decision support systems in clinical scenario[6]. For example, the NegEx[7] is an open source tool that can identify candidate disease names and negated assertions in arbitrary text. Uzuner et al.[8] proposed a statistical system for classifying the assertions in patient health records.

The work reported here presents a hybrid approach for identifying the medical concepts in patient health records. It utilizes a CRF-based supervised classifier combined with the ProMiner system. For the classification of medical assertions, a Support Vector Machine (SVM)-based system was applied. The performances of both the approaches have been demonstrated on the training set as well as on an independent test set.

2. Goals and the Corpus Characteristics

The 2010 dataset provided by I2B2 contains a training set of 349 expert-annotated patient health records (referred to as *I2b2-train*) and a test set of 477 unannotated records (referred to as *I2B2-test*). The *I2B2-train* corpus was annotated for the mentions of medical *problems*, *treatments* and *tests*. The *I2B2-train* corpus contains 30,673 sentences, 260,573 tokens and 27,831 annotated entities altogether wherein 11,967 entities were annotated as *problems*, 8,497 entities as *treatments* and 7,367 entities as *tests*. The *I2B2-test* corpus contains 45,053 sentences and 396,173 tokens. A later supplied gold standard for the *I2B2-test* contains 45,009 annotated

entities wherein 18,550 entities were annotated as *problems*, 13,560 entities as *treatments* and 12,899 entities as *tests*. The aim of concept identification task is to utilize the information from *I2B2-train* corpora in order to automatically tag the mentions of medical *problems*, *treatments* and *tests* in the *I2B2-test* corpus. The expert annotations of the *I2B2-test* (also referred to as gold standard) were made available at the end for assessing the performance of the applied system.

On the other hand, for assertion classification task, only the mentions of medical problems in the *I2B2-train* corpus were categorized into six predefined categories. The mentions of 8,052 problems were categorized as *present*, 2,535 as *absent*, 535 as *possible*, 103 as *conditional*, 651 as *hypothetical*, and 92 as *not associated with the patient*. The gold standard for the *I2B2-test* contained 13,025 problems categorized as *present*, 3,609 as *absent*, 883 as *possible*, 717 as *hypothetical*, 171 as *conditional*, and 145 as *not associated with the patient*. The aim of assertion classification task is to utilize the information from the *I2B2-train* corpus and automatically classify the mentions of medical problems in the *I2B2-test* corpus into pre-defined categories.

3. Methods

3.1 Concept Identification with CRFs

Conditional Random Fields is a machine learning technique that has been widely used for modeling the sequential data. The technical details of CRFs can be found in the report of Roman and Tomanek[9]. The I2B2 data was tokenized at whitespaces and converted into IOB sequences before they can be subjected to training or validation. Table 1 shows an example of text snippet in the IOB format. The labels **B-Prob**, **B-Treat** and **B-Test** indicate beginning tokens of the problems, treatments and tests whereas **I-Prob**, **I-Treat** and **I-Test** correspond to intermediate tokens of problems, treatments and tests respectively. The label **O** corresponds to a token that does not belong to any entity class.

Tokens:	The	rectal	cancer	was	observed
Labels:	**O**	**B-Prob**	**I-Prob**	**O**	**O**

Table 1. Example of text snippet and label sequence after tokenization and IOB conversion.

3.2 Feature Sets used for Concept Identification

The features used for training the CRF can be broadly categorized as *morphological*, *grammatical*, *context-based* and *ProMiner-based* features. *Morphological features* are concerned with the internal structure of the tokens (e.g. suffixes/prefixes, capitalizations, special characters, WordAsClass, etc). *Context-based* features use information about the surrounding elements for every token [e.g. offset conjunction (OC) of order ±1, ±2]. *Grammatical features* are Parts-Of-Speech (POS) tags of the tokens. *ProMiner-based* features include lists of candidate named entities that occur in the complete I2B2 corpus that were recognized by the ProMiner. Three separate dictionaries were used for identifying the candidate names of problems, treatments and tests. For identifying the candidate medical problems, the MedDRA[10] dictionary was used. A combined dictionary composed of entries from DrugBank[11], KEGG[12], Drugs@FDA[13] and UMLS[14] drugs was used for identifying the candidate treatments. A subset of UMLS medical test dictionary was used for identifying the names of tests in the I2B2 dataset.

3.3 Assertion Classification

The principle behind this classification task was to use the contextual information in order to automatically classify the assertions of medical problems. A range of classifiers that include Naïve Bayes[15], Nearest Neighbor[16], Decision Tree[17] and Support Vector Machine[18] (SVM) were preliminarily validated on the training data. Based on the outcome of this validation, the best suited classifier was subjected to classify the instances in the test set. Weka 3.6[19] platform was used for the assertion classification task.

3.4 Feature Sets used for Assertion Classification

During the preliminary evaluation over the training data, various feature sets were tested that include words-in-window, lemmas-in-window, bigrams-in-window, positions, family history and nearest verbs.

a. Words-in-window of size ‘±n’ includes ‘n’ number of words that precede and succeed the mentions of medical problems.

b. Lemma-in-window of size ‘±n’ includes lemmas of ‘n’ number of words that precede and succeed the mentions of medical problems.

c. Bigrams-in-window of size ‘±n’ includes bigrams (also referred to as word pairs) of ‘n’ number of words that precede and succeed the mentions of medical problems.

d. Position adds information to every token, lemma or bigram whether it precedes or succeeds the mention of medical problem.

e. Family history adds information to the mention of medical problem whether it occurs in 'family history' subsection of the document or not.

f. Nearest verbs include the verbs that precede and succeed the mentions of medical problems. The lemmatized forms of verbs were used as features. This feature is independent of the size of the window.

Words-in-window, lemmas-in-window, bigrams-in-window, positions and family history were modeled as binary features. For every feature, its value was set to '1' if the feature was present or set to '0' if the feature was absent. Nearest verbs denoted two separate features i.e. left nearest verb and right nearest verb with their values set to the respective lemmatized stings. Table 2 shows an illustration of features associated with an arbitrary problem concept.

Example: He was noted to have **an erythematous perianal rash** for which he started on Nystatin powder.	
Window size	±4
Words-in-window	was, noted, to, have, for, which, he, started
Lemma-in-window	be, note, to, have, for, which, he, start
Bigrams-in-window	was+noted, noted+to, to+have, for+which, which+he, he+started
Lemma-in-window + Position	PRE=be, PRE=note, PRE=to, PRE=have, POST=for, POST=which, POST=he, POST=start
Nearest verbs	PREVERB=have, POSTVERB=start

Table 2. Features associated with a problem concept *an erythematous perianal rash* that will be subjected to assertion classification.

4. Results and Discussion

4.1 Performance Evaluation Criteria

The evaluation of concept identification was performed using exact match as a criterion. An exact match is a situation where the system identifies both left as well as the right boundaries of the annotated concept correctly. The performances of concept identification and assertion classification were judged based on Precision, Recall and F-score. During the preliminary experiments, the performances of the systems were evaluated by 10-fold cross validation of the *I2B2-train* corpus. Finally, the best performing settings were chosen to tag or classify instances in the *I2B2-test* corpus.

4.2 Evaluation of Concept Identification

Under the preliminary settings, the CRF was trained and evaluated by 10-fold cross validation of the *I2B2-train* corpus. All the morphological features and context-based features (OC = ±1) were used. The system attained the F-score of 0.78 ± 0.04 (also referred to as baseline).

In order to evaluate the impact of different features, a new set of features were added to the preliminary feature set or the existing ones were set to idle and experimented in a systematic way. For every modified feature set, a separate model was trained and evaluated by 10-fold cross validation. The result of feature evaluation is shown in Table 3. Only the features that contributed to an improvement in the performance of baseline result is shown. Table 3 implicitly indicates that the perturbation of preliminary features set does not contribute to the improvement of the baseline result. The POS tags improved the performance by nearly 1% whereas adding the lemmas and ProMiner-based features contributed substantially to the system's performance. Finally, the best model that achieved the F-score of 0.83 ± 0.03 was applied to tag the *I2B2-test* corpus.

Features	F-score
Prel. Features	0.78 ± 0.04
Prel. features + POS	0.79 ± 0.04
Prel. Setting +POS +Lemma	0.81 ± 0.05
Prel. Setting +POS +Lemma +ProMiner-based	**0.83 ± 0.03**

Table 3. Results of the system's performance (F-score) during different stages of feature evaluation experiments. Prel. features indicate all the morphological features and OC = ±1.

4.3 Evaluation of Assertion Classification

In the first step, different classifiers were trained and evaluated by 10-fold cross validation of the *I2B2-train* corpus. Words-in-window of sizes ±2, ±3, ±4, ±5, ±6 were used as features in the preliminary settings. The aim was to choose one best classifier and a suitable window size for further evaluations. Table 4 shows the performance of different classifiers over the varying window sizes during 10-fold cross validation.

Win. size	NB	NN	DT	SVM
±2	0.76±0.04	0.78±0.05	0.82±0.04	0.84±0.05
±3	0.77±0.05	0.79±0.04	0.83±0.05	0.85±0.04
±4	0.78±0.05	0.79±0.05	0.83±0.03	**0.86±0.03**
±5	0.78±0.03	0.79±0.06	0.83±0.06	0.86±0.04
±6	0.77±0.04	0.79±0.05	0.84±0.04	0.86±0.03

Table 4. The performance of different classifiers (F-score) over the varying window sizes during 10-fold cross validation.

Based on the results of classifier and window size selection, the SVM and a window size of ±4 were chosen to be optimum for further experimentation. In the second step, different sets of features were used and the performance of SVM was evaluated. For

every modified feature set, a separate model was trained and evaluated by 10-fold cross validation. The results of feature evaluation for the assertion classification task are shown in Table 5.

Features	F-score
Words-in-window (baseline)	0.86 ± 0.03
Bigrams-in-window	0.86 ± 0.04
Lemma-in-window	0.87 ± 0.04
Lemma-in-window + Positions	0.89 ± 0.05
Lemma-in-window + Positions + Nearest verbs	0.89 ± 0.04
Lemma-in-window + Positions + Nearest verbs + Family History	0.90 ± 0.04
Lemma-in-window + Positions + Family History	**0.90 ± 0.03**

Table 5. Results of the system's performance (F-score) during different stages of feature evaluation experiments for the assertion classification.

The results of feature evaluation indicated that a combination of lemma-in-window, positions and family history coupled with SVM is best suited for classifying the assertions of medical problems in the *I2B2-test* corpus.

4.4 Final Evaluation over the Test Set

For identifying the concepts in *I2B2-test* corpus, a trained CRF that utilizes the optimum feature set mentioned in Section 4.2 was applied. The results of concept identification over an independent test set are shown in Table 6. The applied system achieved an overall F-score of **0.82** for tagging the problems, treatments and tests in the *I2B2-test* corpus.

For classifying the assertions in *I2B2-test* corpus, a trained SVM that utilizes the optimum feature set mentioned in Section 4.3 was applied. The results of assertion classification are shown in Table 7. The applied system achieved an overall F-score of **0.90**.

	P	R	F
Problem	0.84	0.80	0.82
Treatment	0.84	0.77	0.81
Test	0.85	0.80	0.82
Overall	0.84	0.80	**0.82**

Table 6. Assessment of performance of the system for identifying the concepts in test corpus.

	P	R	F
Present	0.92	0.96	0.94
Absent	0.88	0.85	0.87
Possible	0.71	0.47	0.57
Hypothetical	0.73	0.73	0.73
Conditional	0.70	0.18	0.28
Not associated with patient	0.96	0.66	0.78
Overall	0.90	0.90	**0.90**

Table 7. Assessment of performance of the system for classifying the assertions in test corpus.

4.5. Error Analysis

The concepts tagged by the CRF during training as well as in the test set were manually investigated to understand some common sources of errors. Examples of frequent sources of errors include abbreviations such as *CXR* that stands for *chest X-ray* and *IVP* that stands for *Intravenous Pyelogram*. Apart from abbreviations, the descriptive enumerations of medical problems caused substantial problems. Long and descriptive mentions of the medical problems such as *subtle decreased flow signal within the sylvian branches* were not recognized completely by the system. Other sources of errors include nested concepts such as *rupture of liver, left renal vein, pancreas, and transverse mesocolon* and anaphors such as *the following medications.*

A manual inspection of the results of assertion classification indicated several errors. For example, in the sentence *It was felt that his dementing illness and rigidity was most likely due to some type of* ***cortico-basal ganglia degeneration process****, but this was not clarified during this admission* , the medical problem *cortico-basal ganglia degeneration process* that was originally annotated as *possible* was misclassified as *present* by the system. This is because the 4 features in the preceding window could not capture the keyword *likely* which is a critical feature in this scenario for a correct classification. Other examples include mentions of multiple neighboring problems that render them far from their actual context. For example, in the sentence *She currently denies any* ***fever*** , ***chills*** , ***night sweats*** , ***weight change*** , ***blurred vision*** , ***headaches*** , ***nausea*** , ***vomiting*** , ***diarrhea*** , ***constipation*** , ***abdominal pain*** , ***changes in vision*** , ***shortness of breath*** , ***chest pain or pressure*** , *or* ***changes in her bowel habits***, concepts such as *vomiting*, *diarrhea*, *constipation* etc. that belong to the class *hypothetical* were misclassified as *present* since their preceding or succeeding features fail to capture the actual context. In the cases of both concept identification as well as assertion classification, the annotation errors induced by human annotators also contribute to the decline in performance of the system.

5. Conclusion

The approaches introduced for medical concept identification with CRFs and assertion classification with SVM achieved competitive results with F-scores of 0.82 and 0.90 respectively. In case of concept identification, it was shown that the application of ProMiner enabled features to CRFs substantially

contributes to the performance of the system. For classifying the assertions on medical problems, the window-based contextual features in combination with SVM were shown to be successful.

Nevertheless, several strategies have to be tested in order to improve the performances of the applied systems. The dictionaries used for identifying the candidate named entities had a limited coverage. For example, the dictionary used for treatments had a good coverage of chemical and drug names but did not include names of operative procedures, therapies, etc. Manual curation and quality assurance of the terminological resources is necessary. Additional features that could enhance the concept recognition capability of the CRFs should be tested.

For the assertion classification, the feature selection and parameter optimization for SVM should be performed in a systematic way. Additional features can be harvested by applying openly available tools like NegEx or MedLEE that could enhance the performance of the system. The system's generalizability to work on different corpora other than the I2B2 dataset should be tested. As a use case scenario, the developed systems are believed to improve literature-based knowledge discovery and thus support hypothesis generation for advanced health care in the medical arena.

6. Acknowledgements

Thanks to Roman Klinger (SCAI), Heinz Theodor-Mevissen (SCAI) and Shweta Bagewadi (B-IT) for technical support. The CRF toolkit used here was developed by Roman Klinger. This work was supported in part by the B-IT Research School within the NRW State.

References

1. https://www.i2b2.org/NLP/Relations/Documentation.php
2. Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J. ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics. 2005; 6 Suppl 1:S14.
3. Lin D, Kipper-Schuler K, Savova G. Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. Current Trends in Biomedical Natural Language Processing. 2008; pp. 94–95.
4. Osborne JD, Lin S, Zhu L, Kibbe WA. Mining biomedical data using MetaMap Transfer (MMtx) and the Unified Medical Language System (UMLS). Methods Mol Biol. 2007; 408:153-69.
5. Chiang JH, Lin JW, Yang CW. Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE). J Am Med Inform Assoc. 2010 May 1; 17(3):245-52.
6. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005 Mar 9;293(10):1223-38.
7. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct; 34(5):301-10.
8. Uzuner O, Zhang X, Sibanda T. Machine learning and rule-based approaches to assertion classification. J Am Med Inform Assoc. 2009 Jan-Feb; 16(1):109-15.
9. Klinger R, Tomanek K. Classical Probabilistic Models and Conditional Random Fields. Technical Report. Dortmund Uni. of Technology, 2007.
10. Merrill GH. The MedDRA paradox. AMIA Annu Symp Proc. 2008 Nov 6:470-4.
11. Wishart DS. DrugBank and its relevance to pharmacogenomics. Pharmacogenomics. 2008 Aug; 9(8):1155-62.
12. Wixon J, Kell D. The Kyoto encyclopedia of genes and genomes-KEGG. Yeast. 2000 Apr; 17(1):48-55.
13. Drugs@FDA:www.accessdata.fda.gov/scripts/cde/drugsatfda/
14. Merrill GH. Concepts and synonymy in the UMLS Metathesaurus. J Biomed Discov Collab. 2009 Oct 14; 4:7.
15. Rish I. An empirical study of the Naive Bayes classifier. IJCAI-01 workshop on Empirical Methods in AI. 2001.
16. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967; 13(1).
17. Quinlan JR. Induction of Decision Trees. Machine Learning archive. 1986; 1(1):81-106.
18. Vapnik VN. The Nature of Statistical Learning Theory. Springer. 1995.
19. WEKA: http://www.cs.waikato.ac.nz/ml/weka/