TTI's Systems for 2010 i2b2/VA Challenge

Yutaka Sasaki, PhD, Kenta Ishihara, Yukiya Yamamoto, Davy Weissenbacher, PhD
Toyota Technological Institute, Nagoya, Japan

Abstract

In this paper, we describe TTI's systems for the 2010 i2b2/VA Challenge. We participated in the concept extraction and assertion classification tasks. The concept extraction task is to predict three kinds of medical named entities: Test, Problem, and Treatment. The assertion classification task is to identify six kinds of contextual statuses of the medical problem, which consist of Present, Absent, Possible, Hypothetical, Conditional, and Not associated with the patient. In both tasks, we took a fully supervised approach using Conditional Random Fields. Our systems did not rely on any external resources, such as UMLS or WordNet. In the concept and assertion tasks, we achieved F1-scores of 80.19 and 90.85, respectively.

Introduction

Text Mining technologies have high potential in contributing to *Medical Informatics*. However, collection of real natural language data in the medical domain is always a major bottle neck due to privacy issues.

In this respect, the 2010 i2b2/VA challenge is a valuable opportunity to gain an access to real-world textual data that can be used for medical text mining. The i2b2 challenge organizers collected and annotated 349 and 377 discharge summaries for training and test, respectively. The number of documents is relatively small but we understand how difficult to get permission to use real clinical documents for a research purpose, which is an inevitable situation in medical text mining. All our team members passed NIH's Web-based human subject training course and obtained a certificate.

We participated in the concept extraction and assertion classification tasks. Since we are interested in revealing the information content that is contained in the training data, in both tasks, we took a fully supervised approach using *Conditional Random Fields (CRFs)*[1] without using any other tools and resources.

In the following sections, we explain Toyota Technological Institute's concept extraction and assertion classification systems that participated in the i2b2 Challenge.

Challenge Overview

The 2010 i2b2 Challenge was designed to evaluate three aspects of medical information processing that will enable clinical researchers to use existing clinical data for discovery research

The concept extraction task evaluates the ability to predict three kinds of medical named entities: medical *problem* (*e.g.*, diseases, injuries), medical *treatment* (*e.g.*, procedures, interventions, substances to resolve a medical problem) and medical *test* (*e.g.*, procedures and measures to discover or rule out medical problems). Discharge summaries are given in the format of tokenized plain texts.

The assertion classification task evaluates the ability to identify six kinds of contextual statuses of the problem description, which are *Present*, *Absent*, *Possible*, *Hypothetical*, *Conditional*, and *Not associated with the patient*. Gold standard concept annotations and discharge summaries are given.

The relation classification task evaluates the ability to extract relations between medical concepts, in which we did not participate.

The brief progress of TTI team was as follows:

April 15, 2010 Registration

April 28, 2010 All team members completed the NIH Web-based training course "Protecting Human Research Participants".

May 1, 2010 Gained access to sample data

June 17, 2010 Training data released

July 26-27, 2010 Concept and assertion tasks formal runs

Corpus Overview

The training and test corpora are discharge summaries from the following medical institutes:

- Partners HealthCare
- Beth Israel Deaconess Medical Center
- University of Pittsburgh Medical Center.

			Training	Test	Total
#documents			349	377	726
# annotations			27,837	45,009	72,846
	Test		7,369	12,899	20,268
	Treatment		8,500	13,560	22,060
	Problem		11,968	18,550	30,518
		Present	8,052	13,025	21,077
		Absent	2,535	3,609	6,144
		Possible	535	883	1,418
		Hypothetical	651	717	1,368
		Conditional	103	171	274
		Unassociated	92	145	237

Table 1. Corpus statistics relevant to the concept extraction and assertion classification tasks.

All summaries are real-world data but they have been fully de-identified.

Concept, assertion, and relation information are manually annotated. **Table 1** shows statistics of the corpora.[a]

System development

We solved both tasks as sequence labeling problems. During the development phase, we used the following 11 file IDs for system evaluations:

910458031
920798564
959086752
979440029_RWH
915093496_RWH
932057504_DH
965367286_WGH
989519730_WGH
917989835_RWH
950452368
974381789.

The remaining data were used for training during the development phase. For the test phase, we used all the models trained on all the training data. Conditional Random Fields have been used to generate sequence prediction models.

According to concept and assertion annotation files (*i.e.*, *.con and *.ast files), we converted text into CRF training data in which each token is given features and correct sequence labels. We employed the standard IOB2 labeling scheme. That is, the label "B-*category*" is given to the first token of the target sequence, "I-*category*" to each remaining token in the target sequence, and "O" to other tokens.

[a] The numbers are based on our counting, not official figures.

	Training time (sec)
Concept	2162
Assertion	2870

Table 2. Training CPU time spent for training concept and assertion CRF models.

Based on our previous experience in NER[2], in our i2b2 systems, the following primitive features are given to each token:

- Word feature
  - Token itself
- Document structure feature
  - Section title
- Word form features
  Capital letters in a token are normalized to "A", lower case characters are normalized to "a", and digits are replaced by "0". For example, the word form "IL-2" is normalized to "AA-0".
  - First character in the word form
  - Last four characters in the word form
- Prefix features
  - first character
  - first two characters
  - first three characters
  - first four characters
- Suffix features
  - last character
  - last two characters
  - last three characters
  - last four characters
- Concept features
  For the assertion classification task, gold standard concept annotations are given to each token using the same IOB2 labeling scheme.

We applied CRF++ using unigrams, bigrams, and trigrams of above features within a window size of ±2 of the current token.

An example of a template for the feature combinations of primitive feature #1 is as follows:

	F1-score
Concept	77.79
Test	85.55
Treatment	78.08
Problem	73.82
Assertion	95.44
Present	97.15
Absence	95.81
Possible	66.67
Hypothetical	88.00
Conditional	00.00
Unassociated	66.67

Table 3. Development F1-scores of our i2b2 systems.

```
U101:%x[-2,1]
U102:%x[-1,1]
U103:%x[0,1]
U104:%x[1,1]
U105:%x[2,1]
U106:%x[-2,1]/%x[-1,1]
U107:%x[-1,1]/%x[0,1]
U108:%x[0,1]/%x[1,1]
U109:%x[1,1]/%x[2,1]
U110:%x[-2,1]/%x[-1,1]/%x[0,1]
U111:%x[-1,1]/%x[0,1]/%x[1,1]
U112:%x[0,1]/%x[1,1]/%x[2,1]
```

Here, `%x[n,m]` indicates *m*-th feature at relative token position *n*, *e.g.*, `%x[-1,1]` represents first feature in the previous token of the current target token. The slash (/) represents a combination of features. We applied the same feature combinations for all primitive features.

After conducting various attempts to tune parameters of our CRF-based systems using the 11 held-out data, we decided to use CRF++'s parameters `-f 5 -c 4` which mean that the minimum frequency of features is set to 5 and the CRF hyper parameter C is set to 4.

Table 2 shows the training speed of our CRF models for concept extraction and assertion classification. We trained the models using a Linux server with 96GB memory and two Xeon W5590 CPUs (4 cores, 3.33GHz) using CRF++'s `-p 4` option which specifies to train a model using four multi-threads in parallel.

Table 3 shows the development F1-scores evaluated on the 11 held-out data.

	F1-score
Concept	80.19
Test	82.48
Treatment	78.32
Problem	79.94
Assertion	90.85
Present	94.42
Absence	89.09
Possible	52.17
Hypothetical	78.18
Conditional	30.84
Unassociated	72.10

Table 4. Test-run F1-scores of our i2b2 systems.

	Execution time (sec)
Concept	44
Assertion	47

Table 5. Execution CPU time spent for processing 377 test data in concept and assertion tasks.

Test-Run Results

The aim of this section is to demonstrate our concept extraction and assertion classification performances on the test data provided by the i2b2 organizers during the test run period.

Tables 4 shows test set performances of our systems. In general, development and test evaluation show the similar tendency. In concept extraction, *Test* is the easiest category. In assertion classification, *Possible* and *Conditional* are difficult to correctly predict. **Table 5** shows execution CPU times for the test runs. It is fast enough to process each text file in around 120msec.

Tables 6 and **7** show the test set evaluations in concept extraction and assertion classification with different feature settings, respectively. The tables

Feature set	F1-score
word	75.81
+ section	74.83
+ word form	79.98
+ prefix	80.19
+ suffix	80.19

Table 6. Contribution of each feature set in concept extraction performance.

show that the word form and the prefix feature sets mostly contributed to the final performances. The detailed results on the test data are shown in **Tables 8** and **9**.

Related work

The concept extraction task is a variant of the *Named Entity Recognition (NER)* task, which is typically solved as a sequence labeling problem. There have been a lot of studies on biomedical NER. We developed our systems based on our previous biomedical NER system without using dictionaries.

The assertion classification task can be regarded as either a standard classification problem or a sequence labeling problem. We regarded the task as a sequence labeling problem to minimize the development time and maximize the performance in the limited development time.

Chun *et al*.[3] used dictionaries and supervised learning to extract disease names from MEDLINE abstract. In our system, we used training data only.

Conclusions and Remark

This paper has presented TTI's systems that participated in the i2bi2 concept extraction and assertion classification tasks. Since we are interested in investigating information content that is contained in the training data *per se*, we took a fully supervised approach without relying of any external resources. We employed state-of-the-art Machine Learning algorithm CRFs to build statistical models of concept extraction and assertion classification. Evaluations on the gold standard annotations show that in the concept and assertion tasks, we achieved F1-scores of 80.19 and 90.85, respectively

Acknowledgements

This research is partly supported by TTI's H22 Research Funds.

Feature set	F1-score
word	89.62
+ section	89.80
+ word form	89.87
+ prefix	90.81
+ suffix	90.85

Table 7. Contribution of each feature set in assertion classification performance.

References

1. Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In Proceedings of the Eighteenth International Conference on Machine Learning (ICML-2001)*. 2001:282–289.
2. Yutaka Sasaki, Yoshimasa Tsuruoka, John McNaught, Sophia Ananiadou. How to make the most of NE dictionaries in statistical NER. BMC Bioinformatics 2008:9(Suppl 11):S5.
3. Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Kong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, Jun'ichi Tsuji. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Pacific Symposium on Biocomputing, 2006:11:4-15.

Exact span for all concepts together						
	True Positive	False Negative	False Positive	R Value	P Value	F Value
Concept Exact Span	36105	8904	5621	0.8022	0.8653	0.8325
Class Exact Span	34775	10234	6951	0.7726	0.8334	0.8019
Exact span for separate concept classes						
	True Positive	False Negative	False Positive	R Value	P Value	F Value
Exact Span for Problem	14876	3674	2759	0.8019	0.8435	0.8222
Exact Span for Treatment	10593	2967	1558	0.7812	0.8718	0.8240
Exact Span for Test	10636	2263	1304	0.8246	0.8908	0.8564
Exact Span With Matching Class for Problem	14575	3975	3340	0.7857	0.8136	0.7994
Exact Span With Matching Class for Treatment	10009	3551	1990	0.7381	0.8342	0.7832
Exact Span With Matching Class for Test	10191	2708	1621	0.7901	0.8628	0.8248
Inexact span for all concepts together						
	True Positive	False Negative	False Positive	R Value	P Value	F Value
Concept Inexact Span	39424	5585	2302	0.8759	0.9448	0.9091
Class Inexact Span	39424	5585	2302	0.8710	0.9040	0.8872
Inexact span for separate concept classes						
	True Positive	False Negative	False Positive	R Value	P Value	F Value
Inexact Span for Problem	16460	2090	1059	0.8873	0.9396	0.9127
Inexact Span for Treatment	11564	1996	658	0.8528	0.9462	0.8971
Inexact Span for Test	11400	1499	585	0.8838	0.9512	0.9163
Inexact Span With Matching Class for Problem	16071	2479	1844	0.8664	0.8971	0.8814
Inexact Span With Matching Class for Treatme	10839	2721	1160	0.7993	0.9033	0.8482
Inexact Span With Matching Class for Test	10811	2088	1001	0.8381	0.9153	0.8750

Table 8. TTI system's test run scores in the *concept extraction* task. Whereas the scores are not official, we evaluated our system using the evaluation script and the gold standard concept annotations provided by the i2b2 organizers.

Exact span for all assertions						
	True Positive	False Negative	False Positive	R Value	P Value	F Value
Exact Span With Matching Concept	18517	33	16	0.9982	0.9991	0.9987
Exact Span With Matching Assertion	16845	1705	1688	0.9081	0.9089	0.9085
Exact span for separate assertions						
	True Positive	False Negative	False Positive	R Value	P Value	F Value
Present With Exact Span	12998	27	13	0.9979	0.9990	0.9985
Absent With Exact Span	3606	3	2	0.9992	0.9994	0.9993
Possible With Exact Span	883	0	0	1.0000	1.0000	1.0000
Hypothetical With Exact Span	715	2	1	0.9972	0.9986	0.9979
Conditional With Exact Span	170	1	0	0.9942	1.0000	0.9971
Associated With Someone Else With Exact Span	145	0	0	1.0000	1.0000	1.0000
Present With Exact Span And Matching Assertion	12768	257	1251	0.9803	0.9108	0.9442
Absent With Exact Span And Matching Assertion	3111	498	264	0.8620	0.9218	0.8909
Possible With Exact Span And Matching Assertion	342	541	86	0.3873	0.7991	0.5217
Hypothetical With Exact Span And Matching Assertion	507	210	73	0.7071	0.8741	0.7818
Conditional With Exact Span And Matching Assertion	33	138	10	0.1930	0.7674	0.3084
Associated With Someone Else With Exact Span And Matching Assertion	84	61	4	0.5793	0.9545	0.7210

Table 9. TTI system's test run scores in the *assertion classification* task. Whereas the scores are not official, we evaluated our system using the evaluation script and the gold standard assertion annotations provided by the i2b2 organizers.