A hybrid approach to extract structured information from narrative clinical discharge summaries

Eric Chang, PHD[1], Yan Xu, PHD[1,2], Kai Hong[1,3], Jianqiang Dong[4], Zhaoquan Gu[3]

Affiliations of the authors: [1]Microsoft Research Asia, Beijing, China; [2]Department of Biological and Medical Engineering, Beihang University, Beijing, China; [3]Department of Computer Science and Technology, Tsinghua University, Beijing, China; [4]Department of Automation, Tsinghua University, Beijing, China

Correspondence: Eric Chang, PHD, Microsoft Research Asia, 5F, Beijing Sigma Center, No.49 Zhichun Road, Haidian District, Beijing 100190, P.R.China;

Telephone number:+86-10-5917-5430

Fax number:+86-10-8809-9511

e-mail:eric.chang@microsoft.com

Abstract:

Object: A hybrid approach is presented as an application of natural language processing (NLP) to clinical discharge summaries for the i2b2/VA relation challenge. The challenge, focusing on translating narrative text to structured representation in the medical domain, consists of three tasks: (1) extraction of concepts including medical problems, treatments, and tests; (2) classification of assertions made on medical problems; and (3) identification of relations between a certain medical problem and another medical problem, treatment, or test.

Design: The overall hybrid approach consists of the following steps: (1) pre-processing the sentences; (2) marking noun phrases (NPs) and adjective phrases (APs) using the improved SharpNLP tools; (3) extracting the concepts using Conditional Random Fields (CRF); (4) classifying medical problems into different assertion categories based on voting by various classifiers; and (5) identifying different relation categories using normalized sentences and voting by various classifiers.

Measurements: The primary performance metrics of the macro-averaged and micro-averaged precision, recall and F-measure, with exact matching and inexact matching, were measured as evaluated results.

Results: The hybrid system achieved a micro-averaged F-measure of 0.7973 for the concept task, 0.9210 for the assertion task and 0.7012 for the relation task, obtaining XX place for concept extraction, XX place for assertion classification, and XX place for relation identification.

Conclusions: The submitted results show our hybrid approach is feasible when the training data are sufficient, indicating that the combined techniques of machine learning and NLP have a promising future in applications of narrative electronic medical records.

Introduction

The digitization of medical records is a growing and inevitable trend along with the adoption of information technology in modern hospitals around the world. Information extraction, which translates narrative text to structured representation, includes concept extraction, assertion classification and relation identification in electronic medical records based on NLP and machine learning techniques. Information extraction is a thriving field and can help physicians obtain patients' current disease status; quickly look up medical records related to patients' diseases; and make accurate clinical decisions from the relations between medical problems, treatments and tests.

The 2010 i2b2/VA relation challenge[1] in NLP for discharge summaries was to design and evaluate systems on their performance in three tasks: concept extraction, assertion classification, and relation identification. Concepts consist of medical problems, treatments, and tests. Assertions of present, absent, possible, conditional, hypothetical, and not associated with the patient are given to medical problems. The relations contain TrIP, TrWP, TrCP, TrAP, TrNAP, TeRP, TeCP, and PIP. The detailed definition of all the categories in these three tasks can be found in the official guidelines of the competition[1]. The i2b2/VA organizers provided a training set with 349 labeled discharge summaries manually annotated with more than 27,000 concepts, 11,500 assertions and 16,400 relations. The test set contained 477 labeled discharge summaries.

Extracting concepts is the starting point for the second and third tasks, and marking concepts is an indispensable part of concept extraction. However, three reasons prevent existing NLP tools to correctly mark NPs and APs that includes gerund phrases. First, discharge summaries have unique characteristics compared with common texts, such as ungrammatical structures of no predicate, no punctuation marks and some acronyms. The second is ambiguity. Words which are both noun attribute and verb attribute in the real world may only have the meaning of noun attribute in the medical domain. The third is various symbols such as “y/o” or “r/o” whose syntactic structure cannot be identified by NLP tools.

In this work, our contribution is four-fold. First, we introduce a parser based on the SharpNLP[2] parser that is designed for the medical domain. Second, we developed a new concept extraction algorithm including both the “medication” model and the “other” model using CRF[3]. Third, our approach combines various classifiers, such as support vector machine (SVM), boosting, logistic-regression (LR) and multiclass logistic-regression[4] (MLR), with the rule-based method to vote as results. Voting results are superior to the results using any of the classifiers or only the rule-based method. Fourth, we use sentence normalization and Link Grammar[5] syntax analysis as features to classify assertions and identify relations. The normalization of sentences plays an important role in obtaining accurate syntax information from the Link Grammar parser. The experiment results demonstrate that the hybrid approach is appropriate for the i2b2/VA relation tasks.

In this paper, a hybrid approach is presented based on NLP tools and machine learning techniques. Fig. 1 shows an overview of the proposed method, which consists of five steps: (1) pre-processing sentences from unstructured discharge summaries; (2) marking NPs and APs using our improved SharpNLP tools[2]; (3) extracting concepts including medical problems, treatments and tests using the “medication” model and the “other” model based on CRF; (4) classifying assertions in the sentences containing medical problems based on voting by various classifiers and features; and (5) identifying relations between the concepts using normalizing sentences and voting by various classifiers.

Related work

In the medical domain, while there are some previous work on NLP tools and information extraction using machine learning techniques, most NLP tools are built by those outside medical informatics who have little knowledge of the medical lexicon. Szolovits[6] built a medical lexicon into the Link Grammar parser to obtain more accurate syntax information. Tsuruoka[7] et al. developed a POS tagger, called GeniaTagger, specifically designed for biomedical texts. Sibanda[8] presented a Category and Relationship Extractor (CaRE) to extract concepts, classify assertions and identify relations based on SVM and the Link Grammar parser. Jagannathan et al.[9] measured commercial NLP engines for extracting medication information.

Methods

In this section, we describe our approach for the i2b2/VA relation tasks from clinical discharge summaries. The approach consists of the following steps -- (1) pre-processing sentences; (2) marking complete NPs and APs using our improved SharpNLP tools; (3) extracting concepts based on CRF; (4) classifying assertions based on voting by various classifiers and features; and (5) identifying relations using normalizing sentences and voting by various classifiers. Fig. 2 shows the general and detailed flow diagrams of the overall method, respectively.

Pre-processing sentences

This step aims to reformat sentences so that the NLP tools can correctly “read” them. Four types of manipulations are performed: (1) removing symbols, such as “1.”, “1)”, “+”, “?”, or “#”; (2) replacing symbols, e.g. replacing “r/o” with “rule out”; (3) dealing with dots which don’t indicate the end of a sentence, e.g. replacing “b.i.d.” with “bid”; (4) splitting digit and dosage using regular expressions.

Marking NPs and APs

Marking NPs and APs has the top priority of the i2b2/VA relation tasks. NLP tools can be used to perform shallow parsing and Part of Speech (POS) tagging. Shallow parsing generates a file, in which NPs and APs are surrounded by round brackets. Existing NLP tools are not suitable for resolving ungrammatical structured records from the medical domain. This is because these records may contain incomplete subjects, predicates, or objects. We modify the SharpNLP tools[2] to make it more suitable for solving the i2b2/VA relation tasks. The modifications are (1) changing verb attribute. We use POS of training data to keep gerund and past participle of regular verbs. The attributes of gerund and past participle of the remaining verbs are changed to the attributes of noun and adjective respectively; (2) adding attributes of medical acronyms to the dictionary of lexical attributes, such as “prn”, “bid”, “po”, and “x”; (3) marking NPs and APs which are the minimum structure, using innermost round brackets. If NPs and APs follow one prepositional phrase that indicates an organ/body part, the entire structure is marked. An organ/body dictionary is built by using body regions in MESH[10] and body part, organ, and organ component in UMLS[11]; (4) smoothing NPs and APs. Because the syntactic structure from discharge summaries may not match the structure from common texts, noise is often generated in marking NPs and APs. In this case, noise that occurs with the same regularity is removed.

Extracting concepts

Extracting concepts, including medical problems, treatments and tests, is the first task in the i2b2/VA relation challenge. Three models, the “medication”, the “other” and the “total”, are designed to extract the concepts using the CRF technique. However, in each Run, only two models, the “medication” model and either the “other” model or the “total” model, are used.

In discharge summaries, “medication on admission”, which describes every day prescriptions for patients, and “laboratory data” or “physical examination”, which describes test results, have different characteristics. “Medication on admission” doesn’t contain complete subject, predicate and object information, while “laboratory data” or “physical examination” has the structure of [test] [digital]. As described above, two models are built for each Run by using different features and training data. Four steps are described in the following paragraphs.

(1) Determining types and appropriate models

The workflow of determining types and models is illustrated as follows. Firstly, a standard dosage unit dictionary is created by using drug part from UMLS and SNOMED_CT[11]. Secondly, an adaptive dosage unit dictionary is generated by combining the standard dictionary with the test data. If a word from the test data partially covers one from the standard dictionary, the word from the test data is considered as a member of the adaptive dictionary. Thirdly, if a sentence from the test data contains a word from the adaptive dictionary, the sentence will be sent to the "medication" model and vice versa.

(2) Generating corresponding features

Although the "medication" model, the "other" model, and the "total" model use the same CRF technique, the training data and the features are different in the three models. For the training data, while the "medication" model uses sentences that contain the words of dosage units, the "other" model excludes dosage unit sentences from the training data. In contrast, the "total" model just applies all the training data. The features used by the "other" model and the "total" model are the same, while those used by the "medication" model are different.

Building corresponding dictionaries is indispensable when generating features. Next, we will describe the dictionaries that are used in our method:

UMLS dictionaries: our three categories are mapped to semantic types in UMLS according to Sibanda's algorithm[6]. When mapping concepts in semantic types to categories, we do normalization by counting the numbers matched and dividing them by the total concepts of each category. If the concept percentage of a certain type is more than a threshold value (we use 80% in our system), we map this semantic type to the category. Table 1 shows that UMLS types are classified into three categories.

MESH dictionaries: Every phrase in MESH belongs to one semantic type by using letters and numbers. “B” and “C” indicates diseases; “D” indicates drugs and chemicals; and “E” indicates equipment and supplies, which are sorted into medical problems, treatments and tests respectively.

Drug-name dictionary: This dictionary is collected from organic chemical, clinical drug, pharmacologic substance, inorganic chemical and lipid in UMLS, and drug descriptions in SNOMED_CT.

Head noun dictionaries: The phrases of three categories from UMLS dictionaries and of the training data are extracted to their own head noun, which is composed of the three corresponding dictionaries using the head noun algorithm[13].

Four categories of features are used by the “medication” model and described in the following.

Lexical context features: The target itself.

Syntactic context features: The POS and the phrases of NPs and APs.

Ontological features: These features include UMLS-based[11] feature, MESH-based[10] feature, drug-name[10,11] feature and head noun feature. Firstly, we map NPs and APs to UMLS or MESH dictionaries using Sibanda’s algorithm[6]. Secondly, we map a noun/adjective phrase containing the word from the drug-name dictionary. Thirdly, the head noun of each phrase from NPs or APs is extracted by the head noun algorithm and maps into the corresponding dictionary.

Orthographic features (five binary properties): These features are used to check whether one sentence contains a dosage unit or temporal adverb, such as “bid” or “q4h”, or if the word contains numerals before dosage units, and drug names before numerals, respectively. Whether a phrase contains characteristics such as [the phrase] [numeral] [dosage] is also checked. The phrase contains a pair of brackets and drug name is included in the interior and exterior of the bracket.

The features used by the “other” model are also divided into the same four types. Lexical context features, syntactic context features and ontological features are the same with the “medication” model, while orthographic features are different, which consists of seven binary properties (see Table 1).

(2) Extracting concepts and matching corresponding types

CRF is widely used in named entity recognition[14]. We use a linear-chain CRF technique to automatically extract concepts and to match corresponding types when the features are completed.

(3) Obtaining correct boundaries

Correct boundary location is the last step in concept extraction. In pre-processing step, the position of the word sequence in a sentence has been undermined. The longest common substring algorithm (LCS)[15] is presented to obtain correct boundaries. LCS is defined as matching the longest substring in the all substrings between two strings from both X and Z. The problem of finding the longest substring can be solved by dynamic programming and a generalized suffix tree between strings of X and Z. In the suffix tree built by a corresponding matrix, we find the deepest internal nodes which have leaf nodes from all the strings by dynamic programming. X and Z refer to the concept being matched and to the original sentence from clinical discharge summaries. Generally speaking, by using this algorithm, the starting position of the concept and the length of the longest substring in Z from the original sentence can be provided by this algorithm.

Three Runs were submitted as follows: Run1: Standard dosage dictionary + “medication” model + “total” model; Run2: Adaptive dosage dictionary + “medication” model + “other” model; Run3: Adaptive dosage dictionary + “medication” model + “total” model.

Classifying assertions

Given clinical discharge summaries and concepts of medical problems, this step classifies each medical problem into one assertion category. There are six assertion categories: Present, absent, possible, conditional, hypothetical and not associated with the patient. In this task, a hybrid approach combining one rule-based classifier and four statistical classifiers is implemented. The part consists of the following steps: (1) manually generating dictionaries and rule-based patterns using training data; (2) classifying assertions using the rule-based classifier; (3) extracting features by combining the above dictionaries and patterns, along with the results from the rule-based classifier; (4) classifying assertions by statistical classifiers; and (5) obtaining the results using voting.

(1) Generating dictionaries and patterns

Effective vocabulary and patterns can be considered as rules in the rule-based classifier and as features in the statistical classifiers. In this task, the dictionaries of key words/phrases for different assertion categories are generated manually. The dictionaries include: words/phrases preceding a problem that indicates that a concept is absent, possible, conditional or hypothetical; Words/phrases succeeding a problem that indicates that a concept is absent, possible, conditional or hypothetical; and words/phrases that indicates that a concept may belong to not associated with the patient. Secondly, some complicated patterns occurring around one concept that may indicate that the concept belongs to one category. For example, when the word "caused" precedes the concept word, there is high possibility that the problem will be classified to the conditional category.

(2) Classifying assertions by using the rule-based classifier

Our rule-based classifier is based on the context window and the rules for each assertion. For not associated with the patient, the result depends on the corresponding dictionaries and section names, such as Family History. For hypothetical and conditional, the result depends on their corresponding dictionaries and some specific patterns. For example, the phrase "for fear" is a symbol for hypothesis; an occurrence of the phrase "on exertion" is an indication of the conditional type. For possible and absent, the result depends on the dictionaries, patterns and context windows. Some additional rules are used to enhance the performance of absent: (1) no/absent/… + or/and is a feature which can extend the scope of the negative words; (2) some words, such as "but", limit the scope of dictionary patterns. Similarly, these rules can be used to enhance the performance of possible. Each concept is checked according to the order: not associated with the patient -> hypothetical -> conditional -> possible -> absent -> present.

(3) Extracting Features

Our assertion features can be divided into three sets: Lexical context features; syntactic context features and rule-based features. Among the three sets, lexical context features and syntactic context features resemble to the one described by Sibanda[8]. Rule-based features are taken mainly from the results of the rule-based method. For each category, we have a double value to signify our confidence concerning whether a problem belongs to this category, which is then added as features of statistical classifiers.

(4) Classifying assertions by statistical classifiers

Four classifiers, SVM, boosting, binary logistic-regression, and multi-class logistic-regression, combined with the above extracted features, are adopted to classify the assertions. One-against-one voting is used as the multi-class classifier from several binary classifiers of the same type. There are four main reasons to choose these four classifiers in our approach: (1) SVM has been shown to be effective in many machine learning projects, and is also feasible when the training set is small or when the input dimensions are high; (2) by combining many weak-classifiers, boosting can achieve high performance in training data and relatively high performance in testing data; (3) logistic-regression[16] is used extensively in medical field, where it can achieve a high performance even when the features are correlated;

(4) multi logistic-regression is used because it preserves the advantage of logistic-regression and has better performance than just logic-regression in results.

(5) Voting

The final results are classified by voting from the five multi- class classifiers.

Three Runs were submitted. Run1: Voting by five classifiers. If two categories had an equal amount of support, the order of categories from associated to present was used. Run2: Possible, conditional and associated categories used the rule-based classifier; the other categories applied voting and tie-breaking methods similar to Run1. Run3: Similar to Run2, except that we applied the result of SVM when two categories had an equal amount of support.

Identifying relations

Relation identification is our third task, which aims at determining the relation between a certain medical problem and other problem, treatment or test. For the case of between two problems, there are two relations, namely PIP (Problem Indicates Problem) and NonePIP. Between a problem and a test, we define three relations, namely TeRP (Test Reveals Problem), TeCP (Test Conducted on Problem) and NoneTeP. Between a problem and a treatment, we include six categories: TrIP, TrWP, TrAP, TrNAP, TrCP and NoneTrP. Different from the second task, there are three parts as we need to identify three types of relations. A hybrid approach combining four statistical classifiers, using SharpNLP and the Link Grammar parser, is implemented. The approach consists of the following steps: (1) normalizing the sentences containing concepts; (2) extracting features including lexical features and syntactically-informed features; (3) identifying relations by using statistical classifiers; and (4) obtaining the results by voting.

(1) Normalizing sentences

It is essential to normalize sentences containing concepts to accurately identify relations for the third task. The process of normalization is performed in four steps: Each concept is replaced with one word place holder. For the relations of two concepts to be identified, the medical problem, treatment, and test are replaced with a certain word and the rest of concepts are replaced with the other word. It is noted that POS of the word place holder is added to the dictionary of SharpNLP and the Link Grammar parser. Some relations always appear in a certain concept with several concepts of the same category. In extracting the lexical features, stemming[12] is used as a part in normalization of sentences.

(2) Extracting features

Our relation features can be divided into two sets: Lexical context features by the improved SharpNLP tools and syntactic context features by the Link Grammar parser, which resemble to the one used by Sibanda[6]. Please refer to [8] for details.

(3) Identifying relations by statistical classifiers

This is similar to that in the second task except for the rule-based classifier. Please refer to the second task description for details.

(4) Identifying relations by statistical classifiers

The final results are identified by voting using four multi-class classifiers.

Three Runs were submitted. Run1: only the SVM classifier was used. Run2: Voting by four statistical classifiers. If two categories had an equal amount of support, the SVM classifier result was then used. Run3: lexical context features + SVM classifier.

Experiments and results

The training data and testing data for the i2b2/VA relation challenge came from four hospitals. The results from three tasks were submitted. The performance was measured using a set of three standard measures: precision (P), recall (R) and F-Measure (F). The results were micro- and macro-averaged for each of the three tasks considered.

Tables 3 and 4 show all the results and the best results in Run2 regarding concept extraction against the test set. As shown in Table 3, the macro-averaged and micro-averaged are (1) exact for all concepts together; (2) exact for diseases, treatments, and tests separately; (3) inexact for all concepts together, and (4) inexact for diseases, treatments and tests separately. The best corresponding micro-averaged F-measure values were 0.7973, 0.7667, 0.8735 and 0.8346, respectively. The best results about exact micro-averaged for all concepts together are shown in Table 4. The micro-averaged F-measures for medical problem, treatment and test were 0.7908, 0.7922 and 0.8122. Three F-measures were almost identical and close to 0.8. The test one was the highest and the medical problem one was the lowest. Compared to the simplest test expression patterns, the medical problem expression patterns were the most complicated.

Tables 5 and 6 show all the results and the best results in Run3 concerning assertion classification for the second task. Table 5 shows the macro-averaged and micro-averaged. The best results for the exact micro-averaged for each of the assertion types separately are shown in Table 6. The present F-measure was the highest due to a large training data set while the conditional one is the lowest, due to sparse training data and difficult patterns for the ruled-based classifier to adapt to. Overall, our micro-averaged F-measure for the present, absent and hypothetical assertions were almost the same for the test set as they were for the training set, which demonstrated that our system was reasonably robust.

Tables 7 and 8 show all the results and the best results in Run2 regarding the relation identification occurring. Table 7 shows the macro-averaged and micro-averaged from all relation types together and Table 8 shows each of the relation types separately. The best corresponding micro-averaged F-measures were 0.7157 and 0.6354.

Discussion

Marking NPs and APs

From Tables 3 and 4, inexact performance was about 10% higher than exact performance. In the experiments on marking NPs and APs, the coverage of accurate phrases was 90% using training data, with 10% of concepts failing to enter into the classification stage. Accurately making noun and adjective phrases exerted a strong influence on the whole performance. However, there was some difficulty when dealing with noun phrases containing a preposition: Some phrases like “shortness of breath” or “hard of hearing” could not be recognized because the word after the preposition was not a body/organ part.

Concept extraction

Run2 had slightly better performance than Run3, which signified that the “medication” model + “other” model was better than the “medication” model + “total” model. Two reasons might cause this phenomenon: (1) all training data were trained in the “medication” model + “other” model, which doesn’t reduce the performance of the system; (2) in the “total” model, “medication” data sometimes acted as noise points and thus interfered with “other” data. Compared to Run2 and Run3, Run1 had the worst performance. In Run1, the dosage dictionary used was the standard one; some “medication” sentences containing “dosage unit” couldn’t be recognized, indicating that the sentences used an error model to extract concepts.

The experiment showed that features played an important role in improving performance. In ontological features, head noun information is indispensable, as the key to the concepts embodies in the head noun itself, improved the value of performance by 2%-3%. UMLS and HESH features improved value of performance by 5%.

Assertion Classification

Among all these classifiers, the rule-based method and SVM obtained better results; multi logistic-regression seemed to guess conditional more often than other classifiers. Compared to Run1, Run2 used a rule-based classifier for three categories, where we got an increase in conditional and associated but not in possible, mainly because the way of using the context window as double increased the performance of voting. Run3 had a slightly better performance than Run2 because it used SVM as the tie-breaking rule instead of following the order from associated to present, which were to be a good choice.

As for the rule-based value, the result was not binary marked. Instead, a confidence value is computed. For instance, if the context window for absent was 3 and the negative word was 4 words away from the concept word, then we gave this word a value of 0.5. According to our experiments, this method proved to be reasonable. Additionally, we added some special patterns to help us make classifications. For example, if a concept word started with “non”, we looked up a dictionary to check whether it was an illness to decide how to assert it. Another example concerned medicine causing an allergy reaction; it meant that “allergy” must then be asserted as “conditional”. However, we should note that the performance for conditional was lower than the others because of the lack of training data and the difficulty of finding rules.

Relation Identification

Experimental results showed that voting improved the performance. The worst one was achieved in Run3 and showed that syntactically-informed features provided considerable information. Without syntactically-informed features, the performance of the micro-averaged decreases about 3% compared to the best Run. The Link Grammar parser often failed to obtain syntactic information due to a lack of grammatical structure. In our Runs, normalization attempted to “fix” this poor grammatical structure.

Conclusion

The paper represents a realistic clinical application that combines NLP and machine learning techniques. Our hybrid approach can successfully cope with the i2b2/VA relation tasks. The improved SharpNLP tool is used to mark NPs and APs from clinical discharge summaries. The CRF model technique can extract the concepts. The voting methods, by combining various classifiers, are used to classify the assertions and identify the relations. At the same time, the primary performance metrics of P, R and F with exact matching and inexact matching were measured as our evaluated results. The experimental results show that the new method is feasible in the i2b2/VA relation tasks.

We believe that the performance can be further improved, especially the performance in exact matching. One possible way of doing this is to improve the accuracy of marking NPs and APs by means of post-processing the extracted NPs and APs. Another possible way is to accurately quantify NPs and APs containing prepositions. In assertion classification and relation identification aspects, dictionaries and patterns being automatically generated instead of manually generated is a promising area. In order to solve the problem of sparse training data for some categories in assertion and relation, we can combine active-learning with semi-learning to automatically generate results from a large amount of unlabeled data.

Acknowledgement

This work is supported by Microsoft Research Asia (MSRA). We would like to thank the organizers of the i2b2 NLP challenge, especially Dr. Ozlem Uzuner and Dr. Scott Duvall for their tremendous contribution.

Reference

1. i2b2/VA relation challenge. Available at : https://www.i2b2.org/NLP/Relations/.

2. SharpNLP tools. Available at: http://sharpnlp.codeplex.com/.

3. CRF++. Available at: http://crfpp.sourceforge.net/.

4. Statistical Classifiers. Available at: http://research.microsoft.com/en-us/downloads/19f63ff3-06c7-4fa9-8ee0-35abffe0e5be/default.aspx.

5. Link Grammar parser. Available at: http://www.link.cs.cmu.edu/link/.

6. Szolovits P. Adding a medical lexicon to an English parser. AMIA Annu Symp Proc 2003:639-43.

7. Tsuruoka Y, Tateishi Y, Kim J, et al. Developing a robust part-of-speech tagger for biomedical text. Adv Inform 2005:382-92.

8. Sibanda TC. Was the patient cured? Understanding semantic categories and their relationships in patient records. MIT master dissertation 2006.

9. Jagannathan V, Mullett CJ, Arbogast JG, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. International journal of medical informatics 2009(78): 284-291.

10. MESH Knowledge Base. Available at: http://www.ncbi.nlm.nih.gov/mesh.

11. UMLS Knowledge Base. Available at: http://www.nlm.nih.gov/research/umls.

12. Stemmer. Available at: http://tartarus.org/~martin/PorterStemmer/.

13. Head noun finder. Available at: http://nlp.cs.berkeley.edu/Main.html#Parsing.

14. Lafferty J, Mccallum A, Pereira F. Condtional random fields: Probabilistic models for segmenting and labeling sequence data. ICMAL 2001: 282-9.

15. Bergroth L, Hakonen H, Raita T. A survey of longest common subsequence algorithms. Proceedings of the Seventh International symposium on String Processing Information Retrieval 2000: 39-48.

16. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. Journal of clinical epidemiology 2001(54): 979-85.

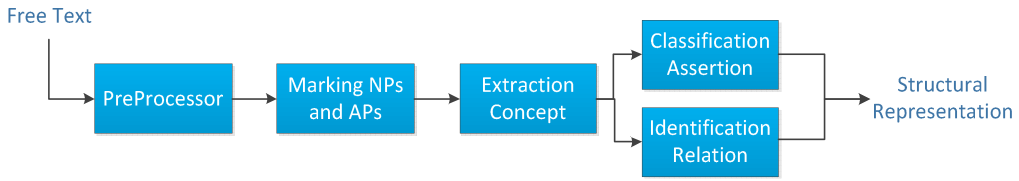

Figure1. The flow diagram of the method

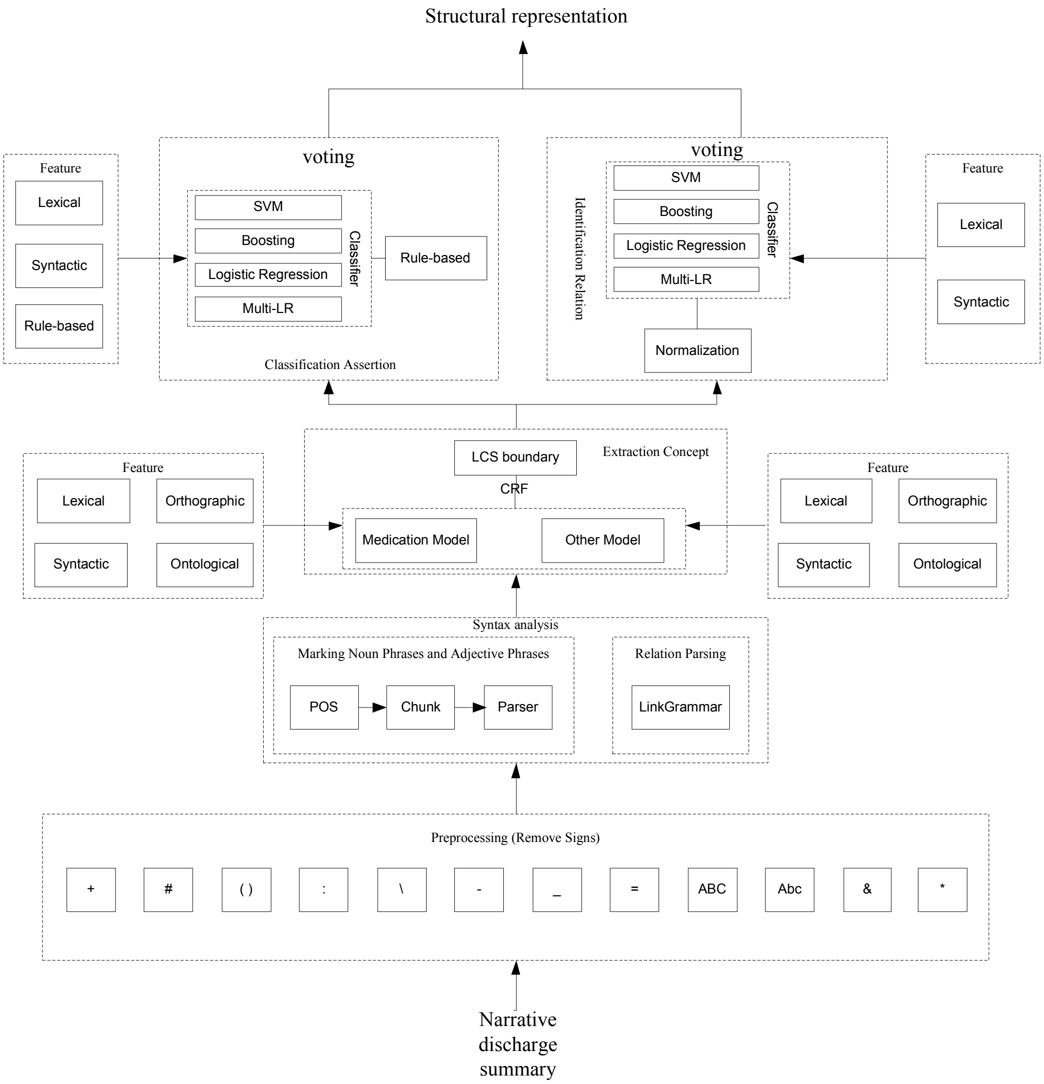

Figure2. The overall flow chart of the method in detail

Table 1 three types of dictionaries from UMLS

concept	Semantic types from UMLS
Medical problems	Finding; Injury or Poisoning; Virus; Disease or Syndrome; Pathologic Function; Bacterium; Organism Function; Sign or Symptom; Neoplastic Process; Acquired Abnormality; Anatomical Abnormality; Mental or Behavioral Dysfunction; Phenomenon or Process; Mental Process
Treatment	Therapeutic or Preventive Procedure; Organic Chemical; Carbohydrate; Hormone; Organophosphorus Compound; Antibiotic
Test	Diagnostic Procedure; Biologically Active Substance; Nucleic Acid, Nucleoside, or Nucleotide; Laboratory Procedure; Clinical Attribute; Research Activity

Table 2 orthographic features in the “other” model for concept extraction

No	Orthographic features
1	Is the word capitalized?
2	Is the entire word capitalized?
3	Is the word abbreviation or acronym?
4	Is the phrase before the numerals?
5	Does the word contain punctuation?
6	Does the word contain assertion word?
7	Does the word contain body/organ word?

Table 3 Macro-Averaged and Micro-Averaged Results for Concept Extraction

	Macro-Averaged			Micro-Averaged		
	P	R	F	P	R	F
Concept Exact Span(Run1)	0.8119	**0.7756**	0.7934	0.8081	**0.7756**	0.7916
Concept Exact Span(Run2)	**0.8245**	0.7741	**0.7985**	**0.8220**	0.7740	**0.7973**
Concept Exact Span(Run3)	0.8244	0.7732	0.7980	**0.8220**	0.7736	0.7971
Class Exact Span(Run1)	0.7750	0.7381	0.7561	0.7713	0.7403	0.7555
Class Exact Span(Run2)	0.7926	**0.7427**	**0.7668**	0.7905	**0.7443**	**0.7667**
Class Exact Span(Run3)	**0.7928**	0.7422	0.7667	**0.7907**	0.7440	**0.7667**
Concept Inexact Span(Run1)	0.8878	**0.8496**	0.8683	0.8868	**0.8512**	0.8687
Concept Inexact Span(Run2)	**0.9007**	0.8464	**0.8727**	**0.9006**	0.8479	**0.8735**
Concept Inexact Span(Run3)	0.9004	0.8453	0.8720	0.9003	0.8473	0.8730
Class Inexact Span(Run1)	0.8400	0.8062	0.8227	0.8415	0.8019	0.8212
Class Inexact Span(Run2)	**0.8607**	**0.8067**	**0.8328**	0.8605	**0.8102**	**0.8346**
Class Inexact Span(Run3)	**0.8607**	0.8060	0.8324	**0.8606**	0.8099	0.8345

Table 4 Micro-averaged Precision, Recall, and F-measure of our best submission for all concepts separately

	P	R	F
Problem	0.8070	0.7752	0.7908
Treatment	0.8251	0.7617	0.7922
Test	0.8411	0.7851	0.8122

Table 5 Macro-Averaged and Micro-Averaged Results for Assertion Classification

	Macro-Averaged			Micro-Averaged		
	P	R	F	P	R	F
Run1	0.7952	0.7056	0.7477	**0.9211**	**0.9211**	**0.9211**
Run2	0.8070	0.7191	0.7605	0.9206	0.9206	0.9206
Run3	**0.8076**	**0.7216**	**0.7622**	0.9210	0.9210	0.9210

Table 6 Micro-averaged Precision, Recall, and F-measure of our best submission for all assertion types separately

	P	R	F
present	0.9347	0.963	0.9490
absent	0.9377	0.9218	0.9297
possible	0.7174	0.5922	0.6488
hypothetical	0.8419	0.7057	0.7678
conditional	0.5272	0.3391	0.4128
associated	0.8863	0.8068	0.8447

Table 7 Macro-Averaged and Micro-Averaged Results for Relation Identification

	Macro-Averaged			Micro-Averaged		
	P	R	F	P	R	F
Concept Exact Span(Run1)	0.7450	0.6502	**0.6944**	0.7694	0.6689	**0.7157**
Concept Exact Span(Run2)	0.6577	**0.6861**	0.6716	0.7192	**0.6841**	0.7012
Concept Exact Span(Run3)	**0.8090**	0.5375	0.6459	**0.8144**	0.5793	0.6770
Matching Relation(Run1)	0.5193	0.4199	0.4643	0.6817	0.5926	0.6340
Matching Relation(Run2)	0.4899	**0.4851**	**0.4875**	0.6516	**0.6198**	**0.6354**
Matching Relation(Run3)	**0.5860**	0.3634	0.4486	**0.7441**	0.5292	0.6185

Table 8 Micro-averaged Precision, Recall, and F-measure of our best submission for all relation types separately

	P	R	F
TrIP	0.7096	0.7777	0.7422
TrWP	0.6947	0.6364	0.6642
TrCP	0.4335	0.7050	0.5369
TrAP	0.7255	0.7821	0.7527
TrNAP	0.5307	0.7696	0.6282
PIP	0.5961	0.3449	0.4370
TeRP	0.8525	0.8197	0.8358
TeCP	0.7191	0.6531	0.6845