Semantic Tagging for Clinical Documents Using a Conditional Random Field

Authors,

Institution

Abstract. This paper proposes a semantic tagging system for clinical documents. The semantic tags in the system are categories of information such as symptom, diagnosis, treatment, etc. Segmented phrases in the document are automatically labeled with these tags. The system uses a Conditional Random Field (CRF) and achieves 80.92% accuracy, which improves upon the baseline by almost 10%. Tf-idf inspired features contributes a significant portion of the increase.

1 Introduction

We propose a semantic tagging system for clinical documents. The tags in this system are categories of information for phrases contained in clinical records such as symptom, diagnosis, and treatment. The system uses a Conditional Random Field (CRF) [1] with the various features in order to annotate the documents.

We use free-style narrative texts of “progress after hospital stay” in the CDA (Clinical Document Architecture) format, provided by XXXX Hospital for research purposes. The texts contain treatment history of individual patients.

The texts have characteristics typical of electronic health records as they are written in free-memo style with frequent use of abbreviations, symbols, numeric expressions, and medical terms [2][3], but are unique because the documents are written in mixed Korean and English. This makes linguistic analysis unreliable, and our best system does not rely on POS taggers or parsers. A set of features inspired by tf-idf, which do not utilize any linguistic information, give the largest gains in the CRF over the HMM baseline.

2 Relation to Prior Work

The 2010 Integrating Informatics with Biology and the Bedside (i2b2)/Veteran’s Affairs (VA) Shared Task in Natural Language Processing for Clinical Data includes a task on concept extraction for noun and adjective phrases in clinical records [4]. The labels are one of four categories: medical problem, treatment, test, and none. Semantic tagging on medical terms using the Unified Medical Language System (UMLS) was done in [5]. [6] describes a medical information extraction system from
free-text clinical records. An overview of information extraction in electronic health records is given in [7].

[8] Presents a semantic tagger which extracts three types of information, “symptom”, “therapy”, and “performance” from free-text clinical records written in mixed Korean and English. The system uses a Hidden Markov Model (HMM) with equivalence classes based on primitive tagging results such as UMLS and POS to solve the sparse data problem.

3 Semantic Tagging Using CRF

In this section, we describe a semantic tagging system for clinical documents. The system uses a Conditional Random Field (CRF) [1]. Tagging is performed at the phrase-level, and the tags in this system are categories of information that phrases of clinical records contain, such as symptom, diagnosis, treatment, and performance.

3.1 Semantic Target Tag

We chose the semantic tags to accurately represent the range of semantic content contained in clinical documents, and to be of interest to health care professionals. The tag set was chosen with the help of a medical professional. The tagging system assigns the appropriate tag to represent the semantics of a given phrase.

For representing the semantics of phrases, we chose the 10 tags described as follows. **Symptom** denotes a phrase containing the state of a patient, and **Diagnosis** denotes a phrase containing the medical opinion of a doctor. **Test** indicates a medical examination, and **Test Result** is the result of the examination. **Treatment** is anything a medical expert performs for the improvement of the patient’s status, i.e. an injection, operation, etc. **Performance** is the result of the treatment, indicating improvement or worsening of the patient. **Treatment Stop** is for a phrase indicating a doctor decided to stop a certain treatment. **Patient Result** is the result of a hospital stay, which is usually ‘discharged from hospital’ or ‘died’. **Garbage** is for phrases not in above categories.

3.2 Features for CRF

In order to use a CRF, features are extracted from the documents. The features can be divided into syntactic and lexical features, and semantic features.

**Syntactic and Lexical Features:** The syntactic and lexical features are tokens, position of a phrase in the document, and Part of Speech (POS) tags of each word. The results of morphological analysis are used for Korean words. POS tagging has been done automatically both for Korean and English.

**Semantic Features:** Semantic features are UMLS tags, clue words, numeric expressions, and symbols. UMLS tagging has been done automatically using MetaMap Transfer (MMTx). Clue words are the key words which frequently appear
within a certain target semantic tag and decide the semantic role of the phrase. For example, if "발열" (having fever) appears in a phrase, it is highly possible that the phrase is symptom.

UMLS tags and clue words are grouped manually as well as automatically. Automatic grouping is according to a method similar to tf-idf weight (term frequency-inverse document frequency). Instead of document in the original tf-idf, we use tag class, and we call this tf-icf (term frequency – inverse class frequency). The tf-icf is calculated for each UMLS tag and token that exists in the documents using the formula shown below.

\[
\begin{align*}
tf & = \frac{\text{# of occurrences of term in a class}}{\text{# of occurrences of term in any class}}, \\
\text{icf} & = \log \frac{\text{total # of classes}}{\text{# of classes containing term}}
\end{align*}
\]

Then, the best UMLS tags and clue words have been selected for a tag class whose tf-icf is over a certain threshold. The threshold is determined by experiments. The tf-icf is suggested for obtaining good features that appear frequently and uniquely with a certain target semantic tag.

4 Evaluation

We compare the performance of our system to a HMM system on the task of semantic tagging in clinical texts. We used 598 narratives from the section “progress after hospital stay” provided by XXXX Hospital. The corpus consists of 7976 phrases allowing nested tags, however only outer tags are considered for the current research. Phrase boundary detection was not performed in the current research. The level of accuracy is computed as the number of correct tags per the total number of tags.

Figure 1 represents the comparison between the performances of the baseline system using HMM [8] and the performance of the current system using various combinations of features. We used 5-fold cross-validation. Note the baseline HMM system was evaluated on a subset of our dataset with a coarser tag set consisting of only Treatment, Symptom, Performance, and Garbage. The CRF system gives the best accuracy of 80.82% when using the feature combination of Token, Numeric expression, Clue words grouped manually, Symbol, UMLS tags, UMLS tags grouped automatically, and Clue words grouped automatically, which is around 10% improvement compared with the best accuracy of the baseline system using HMM.

5 Conclusion

We showed a semantic tagging system for clinical documents achieving 80.82% accuracy, improving upon the baseline. The result shows the features obtained from modified tf*idf have a greatly positive effect while POS tags do not have a positive effect with many other features. For future work, we plan to improve the performance
of primitive tagging such as UMLS tagging considering characteristics of mixed Korean and English.

![Figure 1](image)

**Fig. 1.** Performance comparison on various feature combinations to HMM baseline. The features are T: token, U: UMLS tags, UF: UMLS tags grouped automatically, C: clue words grouped manually, CF: clue words grouped automatically, N: numeric expression, S: symbol, Position: position of a phrase in a document, POS: part of speech tag

### References