PROCEEDINGS OF SCIENCE

An Intelligent Writing Assistant Module for Narrative Clinical Records based on Named Entity Recognition and Similarity Computation

Tianshu Zhou[1,2,3]
EMR and Intelligent Expert System Engineering Research Center of Ministry of Education, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, College of Biomedical Engineering and Instrument Science, Zhejiang University
Hangzhou, 310027, China
E-mail: zts@zju.edu.cn

Jingsong Li
EMR and Intelligent Expert System Engineering Research Center of Ministry of Education, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, College of Biomedical Engineering and Instrument Science, Zhejiang University
Hangzhou, 310027, China
E-mail: ljs@zju.edu.cn

Inpatient medical records which contain clinical narrative information generated from medical procedures have rich content and unlimited expression capabilities. In this paper, we proposed a novel method to assist health practitioners to write narrative clinical texts through a more efficient and safer manner. The core technologies supporting this work are named entity recognition (NER) and similarity computation: the CRF-based NER in this work has a good performance whose F-score has reached 89.23%, and the LDA model and similarity test has reached a precision of 71.28%. After these fundamental work, we designed and developed an intelligent writing assistant module: at sentence level, we used a conditional random field (CRF) method to train a NER model. When doctors type in an entity, several input candidates will pop up for selection; at paragraph level, we used a Gibbs-LDA++ tool and named entities to characterize the topics and key entities of existing records. When doctors create a new clinical text, the patient's structured data will be used as input to match similar paragraphs. As doctors keep typing in, the matching paragraphs also might change dynamically according to the input content.

ISCC2017
16-17 December 2017
Guangzhou, China

[1]Speaker
[2]This work was supported by Chinese National High-tech R&D Program (No.2015AA020109), National key research and development program (No.2016YFF0103200), and the Fundamental Research Funds for the Central Universities of China.
[3]Corresponding Author

PoS(ISCC 2017)041

1.Introduction

Inpatient medical records which are important parts of Electronic Health Records (EHR) contain clinical narrative information generated from medical procedures in hospital. These free-text documents have rich content and unlimited expression capabilities. The information and knowledge implied by which are very useful and important for the proceeding treatment and secondary use of data such as health text mining [1-2]. Thus, the quality and efficiency of writing free-text medical records for health practitioners is of great importance [3].

Currently, there are several ways of writing narrative medical records in EHR systems: 1) writing from sketch all by hand typing, 2) use auto complete assistant to speed up typing, 3) choose and modify some existing templates [4]. The latter two methods have improved the efficiency of recording clinical narratives a lot, but there are limitations: in most cases, the auto complete assistant is based on some fixed recommendation models and custom dictionaries, thus the completion options will not change according to the situations of a specific patient or the previous input content; as of using templates, health practitioners have to save and organize a group of templates and choose one by category or tags which could easily lead to some mistakes when they do not go through and rewrite the template carefully, and also may result in a bunch of similar clinical documents; the limitation list could go on.

We designed and developed an intelligent writing assistant module for narrative clinical records based on named entity recognition (NER) and latent dirichlet allocation (LDA) to enhance the system operability and data accessibility when writing free-text records. In detail, at sentence level, we used a conditional random field (CRF) method to train a NER model, then we could constantly abstract named entity-centered sentences from growing narrative medical records and store them as templates in a data collection. When doctors type in an entity, several input candidates will pop up for selection, and some values specific to the patient (such as blood pressure: 83/128mmHg) in the candidate sentence will be in a selectable text field whose options are retrieved from structured data in EHR; at paragraph level, we used a Gibbs-LDA++ tool and named entities to characterize the topics and key entities of existing records. When doctors create a new narrative document for a patient, the patient's structured data and other documents will be used as input to match similar paragraphs. As doctors keep typing in, the matching paragraphs may change dynamically according to the input content, doctors could select a paragraph and continue to edit.

2.Method

2.1 Named Entity Recognition

Named entities are key fundamental elements in paragraphs and notes [5]. In clinical narratives, named entities are meaningful phrases such as diseases, symptoms, tests, medications, treatments and so on [6], which constitute the primary lines of the document and could be quickly parsed and reused to support writing assistant. In this work, we used the CRF++ tool to obtain the entities. The pre-work required would be described in the following sections.

2.1.1 Feature Selection

Feature selection is the basis of preparing corpus and feature templates when using CRF++ tool, which also plays an important role to balance the accuracy and efficiency to recognize named entities. In this work, we considered four types of feature:

(1) Context window length

In Chinese medical records, the disease name length is about 2 to 7 characters, clinical manifestation about 2 to 11 characters, and surgical operation about 5.2 characters[1]. We chose a context window length of 7, which could cover most named entities in this study and would be efficient as well to perform the training process.

(2) Token types

Chinese linguistic token types related to this study are characters and words, which both could be selected as the feature. Considering that the word segmentation would complicate our work and might even make the NER result worse, we dropped that option and chose single character as the feature.

(3) Part of speech

There are two categories and 12 types of part of speech in modern Chinese. In general, named entities are noun. In many cases, verbs such as "suffer", "inject" and "take" are often followed up with entities like disease name and medication in the clinical text. Thus, part of speech could be a feature to distinguish the named entities from other words.

(4) Formation patterns

There are patterns in Chinese medical terminologies. For example, a Chinese disease entity could be consisted as decoration, common names, body parts and common disease names such as "Alpha thalassemia" in Chinese shape. We sorted out a glossary from the corpus as Table 1 to represent the entity formation patterns.

marker	implication	example
CDN	Common Disease Name	tuberculosis
DE	Decoration	chronic
BP	Body Part	upper limbs
OA	Operation Action	traction
CN	Common Name	Alzheimer
IW	Indicator Words	suffer
NW	Negative Words	not

Table 1： Entity formation pattern features

2.1.2 Corpus and Tagging

The corpus is a set of tagged clinical narrative records for CRF training and testing. In this work, 650 discharge summaries are collected from some hospitals during September 2013 through December 2013 in Zhejiang, China. We need to tag 5 types of named entity which almost cover all the key information in a narrative medical record, as shown in Table 2.

Marker	Implication	Example
DI	disease	Parkinson's Disease
TE	test	troponin
ME	medication	aspirin
TR	treatments	bone tumor resection
CM	clinical manifestation	right lower extremity edema

Table 2： Named Entity Types

We also used a BIEO encoding pattern to tag the border of named entities as shown in Table 3. The marker is consisted of BIEO sign and named entity type, where "B" means left

border of a named entity, “E” right border, “I” internal part of a named entity, and “O” not in a named entity.

marker	implication	marker	implication
B-DI	left border of disease	I-DI	internal of disease
E-DI	right border of disease	B-TE	left border of test
I-TE	internal of test	E-TE	right border of test
B-ME	left border of medication	I-ME	internal of medication
E-ME	right border of medication	B-TR	left border of treatments
I-TR	internal of test treatments	E-TR	right border of treatments
B-CM	left border of clinical manifestation	I-CM	internal of clinical manifestation
E-CM	right border of clinical manifestation	O	none named entity

Table 3： Named Entity Tag Set

Figure 1 is an example of tagged Chinese clinical text. Sentences are separated by one blank line. The first column is a single character, the second part of speech, the third entity formation pattern tag, the fourth indicator word, and the last column tagged entity.

3	m	O	O	O
years	q	O	O	O
ago	f	O	O	O
exams	n	O	O	O
showed	v	O	O	O
that	p	O	O	O
patient	n	O	O	O
had	v	O	B-IW	O
anal	n	B-BP	O	B-DI
polyps	n	B-CDN	O	I-DI
,	wd	O	O	O
and	c	O	O	O
had	v	O	B-IW	O
a	p	O	O	O
anal	n	B-BP	O	B-TR
polyps	n	B-CDN	O	I-TR
resection	v	B-OA	O	I-TR
surgery	n	E-OA	O	E-TR
.	wj	O	O	O

Figure 1.: An example of tagged text

```
#Unigram
U01:%x[-3,0]
U02:%x[-2,0]
U03:%x[-1,0]
U04:%x[0,0]
U05:%x[1,0]
U06:%x[2,0]
U07:%x[3,0]

U11:%x[-3,1]/%x[-2,1]
U12:%x[-2,1]/%x[-1,1]
U13:%x[-1,1]/%x[0,1]

U21:%x[-3,2]/%x[-2,2]
U22:%x[-2,2]/%x[-1,2]
U23:%x[-1,2]/%x[0,2]

U31:%x[-1,3]
```

Figure 2.： A feature template instance

After feature selection and corpus tagging are done, we could prepare several feature templates for NER training. A template example is shown in Figure 2. The CRF++ tool could train corpus with these templates and each template would output a unique model all based on the same corpus. Then, we could use the models to test and implement the named entity recognition.

2.2 Similar Paragraph Recommendation

Similar paragraph recommendation is also based on NER, we would use named entities as input and a LDA model to compute the semantic similarity between clinical texts. The named entities are iterated by a Gibbs sampler, which would produce the optimal number of topics and thus the LDA model to form the semantic space. When users create a new clinical text, the writing assistant system would first fetch the structured data of that patient and compute the semantic similarity within other clinical texts in the semantic space model. After that, three medical narratives sorted by similarity value will be listed. When users keep typing in, the

entities in the content would also take part in the semantic similarity computation, and the listed medical narratives will change dynamically and repeatedly.

3.Experiment and Results

We carried out some tests for NER and similarity recommendation to verify the writing assistant module. There are totally 650 selected discharge summaries, 70% for training and 30% for testing. The count shows that there are 29376 sentences and 425017 characters. The proportion of each type of entities in 650 notes is 35.13% for symptom, 17.17% for medication, 32.26% for test, 9.18% for treatment, and 6.26% for disease.

3.1 NER Testing

We used 4 templates to train the CRF models. The feature combinations are shown in Table 4. Token (Chinese character) and context window length are the basic features required in template. Part of speech, formation patterns and indicator words are optional features which may enhance the precision and reduce the efficiency of NER.

template	feature combinations
T1	Token + context window length (CWL) + part of speech (PS)
T2	Token + CWL + PS + formation patterns (FP)
T3	Token + CWL + PS + indicator words (IW)
T4	Token + CWL + PS + FP + IW

Table 4: Template Feature Combinations

Recall, precision and F-score are used to measure the NER testing results as shown in Figure 3. The best model is trained with template T4. An average recall of 5 types of named entity is 89.84%, precision 88.63% and F-score 89.23%, which could basically fulfil the requirements of writing assistant usage.

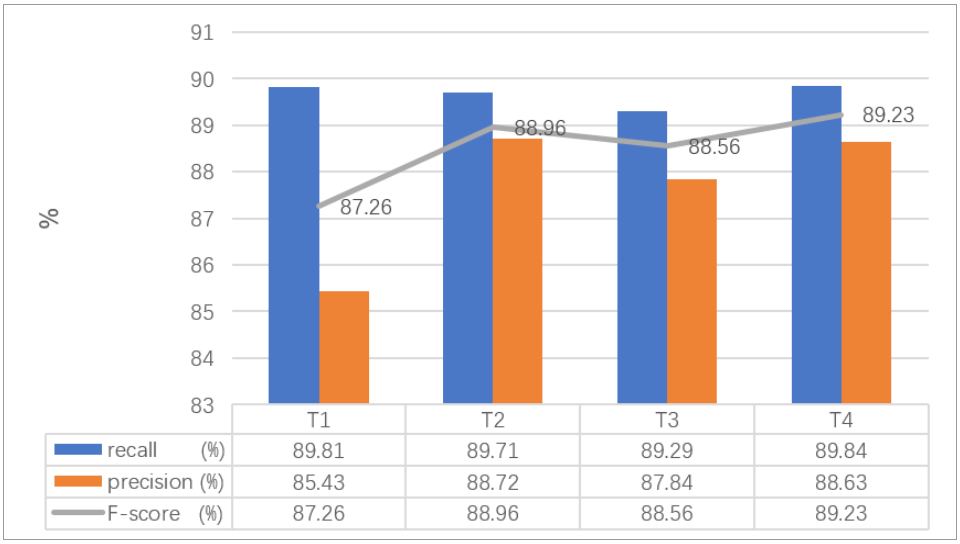

Figure 3:Testing results

3.2 Similarity Recommendation Testing

In this work, we used Gibbs-LDA++ tool to train the LDA model and test the similarity result. The corpus and data partition are the same as mentioned above. The steps are as follows:

(1) Use NER tool to obtain named entities from patients' clinical text, then construct word vector space of the corpus;

(2) Train semantic space using Gibbs sampling method, set $\alpha=50/K$, $\beta=0.01$, iterate 1000 times and sample every 100 iterations. Compute the derivative of perplexity to determine the

optimal topic number K; perplexity is the uncertainty to determine a document which topic a document belongs to. The formula is given below, where N is the number of named entities in the word vector space and p(word) is the probability of word in corpus:

$$perplexity = e^{-\sum \log(p(word))/N} \tag{3.1}$$

(3) Compute similarity when a new clinical text is created using structured data in the patient's EHR and input content; the parameters in the inf file of Gibbs-LDA++ and predictive likelihood method are used to do the job. The formula is given below, where $p(w_m|w_q)$ represents the possibility of document w_m in corpus derived from query document w_q and K is the latent semantic topic number. z means latent semantic of whom w_m belongs to:

$$p(w_m|w_q) = \sum p(w_m|z=k)\, p(z=k|w_q) \tag{3.2}$$

When the optimal number of topics is 40, the perplexity is at its minimal value, and the average similarity precision is 71.28% (by measuring the named entities matching proportion).

3.3Writing Assistant Implementation

The writing assistant system is an independent module which could be integrated into other EHR systems. The user interface is shown in Figure 4. There is a tool bar on the top. The leftmost of the screen is some text fields for typing, and the right side of the screen is a tab menu which could be switched between entity-based sentence recommendation or paragraph recommendation.

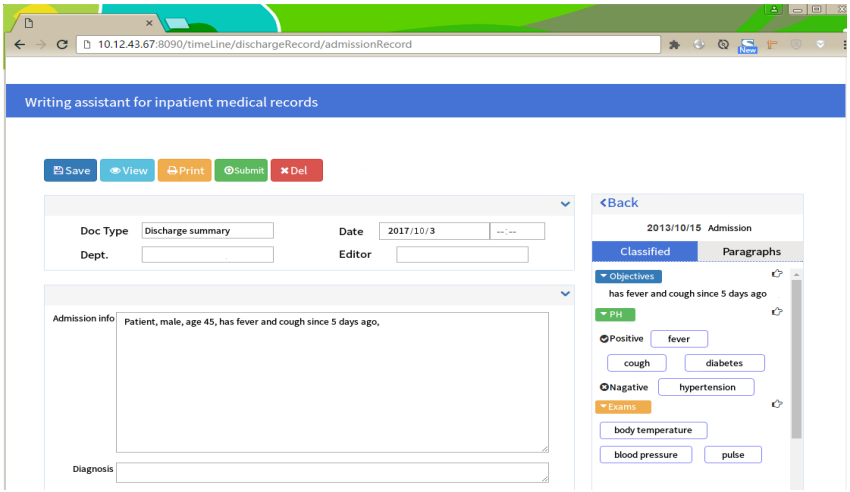

Figure 4: Writing Assistant User Interface

User could drag and drop a similar note recommended on the right side to the text field, and could type in an entity to have several sentence candidates on the right side to select and insert into the text field which is on focus.

4.Discussion and Conclusion

In this paper, we present a novel method to assist health practitioners to write narrative clinical texts through a more efficient and safer manner. The work is mostly accomplished based on named entity recognition and similarity computation, which not only enhance the accessibility and operability of existing structured and free-text data in EHR, but also improve the efficiency of clinical text writing and avoid transcription errors and information missing in some degree compared with the conventional writing mode [7].

The application of CRF-based NER has a good performance whose F-score has reached 89.23% after being trained with a custom feature template. In addition, the LDA model and

similarity test have reached a precision of 71.28%, which could save time for doctors who need to search or find the previous medical records in a paragraph granularity. Overall, the method we propose could increase health practitioners' productivity and improve the quality of narrative medical records.

References

[1] McMillan TE, Allan W, Black PN (2006) *Accuracy of information on medicines in hospital discharge summaries.* Internal Medicine Journal. 36(4):221-225. doi:10.1111/j.1445-5994.2006.01028.x.

[2] Yu-Chia Chang, Jason S. Chang,Hao-Jan Chen, Hsien-Chin Liou. *An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology.* Computer Assisted Language Learning. 2008, 21(3): 283-299.

[3] Shan Li, Tian-Shu Zhou, Xin-Hang Li, Yue-Wen Tu, Jing-Song Li. *The Development of Personalized Writing Assistant for Electronic Discharge Summaries Based on Named Entity Recognition.* ITME 2015 the 7th International Conference on IT in Medicine and Education 660-663,2015.

[4] Wilcox L, Lu J, Lai J, Feiner S, Jordan D. ActiveNotes: *computer-assisted creation of patient progress notes.* Proceedings of the 27th international conference extended abstracts on Human factors in computing systems. New York: ACM, 2009,3323–3328.

[5] Lei JB, Tang BZ, Lu XQ, Gao KH, Jiang M, Xu H (2014) *A comprehensive study of named entity recognition in Chinese clinical text*. Journal of the American Medical Informatics Association. 21(5):808-814. doi:10.1136/amiajnl-2013-002381.

[6] Wilcox L, Lu J, Lai J, Feiner S, Jordan D. *Physician-driven management of patient progress notes in an intensive care unit.* Proceedings of the 28th international conference on Human factors in computing systems. New York: ACM, 2010, 1879.

[7] Schabetsberger T, Ammenwerth E, Andreatta S, et al. *From a paper-based transmission of discharge summaries to electronic communication in health care regions [J]*. International Journal of Medical Informatics, 2006, 75(3-4): 209-215.