Patient Status Classification by using Rule based Sentence Extraction and BM25-kNN based Classifier

Eiji ARAMAKI, Ph.D.[1], Takeshi IMAI, Ph.D.[1]
Kengo MIYO, Ph.D.[1], Kazuhiko OHE, Ph.D., M.D.[1]
[1] The University of Tokyo Hospital, Japan
aramaki@hcc.h.u-tokyo.ac.jp

Abstract *A method for classifying the status of a patient in a medical record is highly desired because this enables larger-scale statistical medical studies. The present paper introduces a system that classifies the smoking status a patient from a medical record. The system consists of two modules: (1) a heuristic-based information extraction module and (2) an Okapi-BM25 and K-Nearest Neighbor-based (kNN-based) classifier module. In experiments, the proposed system achieved an accuracy of 88.97%, demonstrating the basic feasibility of the approach proposed herein.*

Introduction

Medical records contain various types of information that is helpful for statistical medical studies. Automatic analysis remains difficult, however, because most records are written in natural language. The present paper introduces a system that classifies the smoking status of a patient in a medical record.

In this challenge, the smoking status of a patient is categorized into five types as follows:

(C) Current Smoker,

(P) Past Smoker,

(S) Smoker,

(N) Non-Smoker,

(U) Unknown.

An example of a medical record is shown in Figure 1. The red underlined text indicates a sentence that refers to the smoking status of a patient. As shown in the figure, a few sentences (usually only one sentence) refer to the smoking status of a patient.

Therefore, we decompose this task into two processes as follows:

```
PAST MEDICAL HISTORY :
Cryptogenic cirrhosis with an unclear work up .
Diverticular bleed requiring colostomy on May 96 .
Spinal stenosis .
Hysterectomy at the age of 32 .
No coronary disease , no diabetes and no hypertension .
She had medications on transfer which included Synthroid 0.15 mg
 QD , Zoloft 50 mg QD , Inderal 10 mg BID , Prilosec 20 mg QD .
      :

SOCIAL HISTORY :
She does not smoke tobacco .
She uses only occasional alcohol and she is not sexually active .
Her husband died of lung cancer .
She lives alone in Burg Chi Sternafre , Massachusetts .
Her friends check up on her .
She does not have a home health aide .
She has 3 kids who live far away .
      :
```

Figure 1: An Example of a Medical Record.

1. **Smoking status sentence extraction**: First, the system extracts a sentence which refers to the patient smoking status. In this paper, we call the extracted sentence a **Smoking Status Sentence** (shortly S^3). In the S^3 extraction, we use a set of keywords(such as “smoke”, “tobacco” and so on) and heuristic rules.

2. **Classification by using S^3**: The system then classifies the record based on the similarity between S^3 from the input record and S^3s from the training-set records. In this classification, we use Okapi-BM25[1, 2] or the K-Nearest Neighbor (KNN) Classifier[3, 4], both of which are state-of-the-art document classification methods.

Related Work

literature. However, if we regard this problem as a combination of two NLP techniques, namely information extraction and document classification, several previous studies appear in the literature.

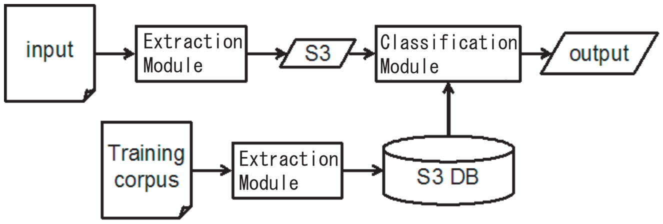

Figure 2: System Workflow.

Information Extraction

Several IE applications, including resume IE[5], seminar announcement IE [6], job posting IE [7, 8] and address segmentation [9, 10], have been reported. While most of these approaches extract information directly from the texts, a few approaches employ two or more steps to extract information. For example, Sitter and Daelemans [7] proposed a two-stage extraction method that works by extracting words from pre-extracted sentences. Their approach is instructive for understanding the approach proposed herein, which uses pre-extracted sentence (S^3).

Document Classification

Because the document classification is a traditional task in the natural language processing field, many methods are proposed.
Document classification is one of the most traditional tasks in the field of NLP and remains an active research area, with several workshops, such as TREC[1] and NTCIR[11] being held recently. The biggest difference in the classification task for this challenge is that this task is sensitive to only a few words. For example, given a text that includes "***no*** *smoking*", only this phrase (especially the word "***no***" is important, and the other sentences are unrelated. Therefore, as mentioned earlier, we first extracted the most important words and then the applied the classification technique[12], KNN[3, 4] and BM25[2], which demonstrated the highest accuracy in the NTCIR patient classification task[11].

Method

The workflow of the proposed system is shown in Figure 2. As shown in the figure, the proposed system consists of two modules.

(1) Information Extraction Module

First, the system extracts a smoking status sentence (S^3), which describes the patient smoking status. The extraction involves the use of a set of keywords: "*nicotine, smoker, smoke, smoking, tobacco, cigarette*", and regards sentences that include any of these keywords as S^3s.
If there are two or more S^3s, we regard the last one as a S^3.
If no S^3 are found in a record, the system classifies the smoking status as UNKNOWN.
Although the proposed extraction method is based on a simple heuristic, it can provide a clear boundary between C,P,S,N records and U records. Table 1 shows the ratio, which is defined as follows:

$$\frac{\#\text{ of records that include } S^3}{\#\text{of records}}.$$

Table 1: S^3 Extraction Ratio.

Smoking Status	Ratio
UNKNOWN	1.1% (= 3/252)
not UNKNOWN (C,P,S,N)	98.6% (=144/146)

As shown in the table, the system usually extracts S^3 from C,P,S,N records (98.6%), but not from U records (1.1%). A number of S^3 examples are shown in Table 2.

(2) Classification Module

The classification module classifies a record based on Okapi-BM25 similarity[2] and K-Nearest Neighbor(kNN) classifier[3, 4].
First, the system calculates the similarity (sim_{BM25}) between the S^3 obtained from an input record (S^3_i) and the S^3 obtained from a training-set (S^3_t). The similarity is defined in Table 3 (for details, see [2]).
The system then extracts the highest similarity k records from the training-set, and the smoking status is decided by the sum of their similarities, as follows:

$$\sum_{S^3_t \in S} sim_{BM25}(S^3_i, S^3_t), \qquad (1)$$

where S is a set of S^3_ts that shares the same status.

Experiments

Experimental Setting

We used a corpus that is provided in the i2b2-NLP shared-task. The corpus consists of 398 records and their smoking status tags.
The number of each tag is shown in Table 4. By five-fold cross validation, we compared the following three methods:

[1] http://trec.nist.gov

Table 2: Examples of Smoking Status Sentences (S^3s). A bold word indicates a keyword.

Smoking Status	Smoking Status Sentence (S^3)
NON-SMOKER	She does not **smoke tobacco** .
NON-SMOKER	The patient does not **smoke** .
NON-SMOKER	He does not drink alcohol , **smoke** or use illicit drugs .
SMOKER	PAST MEDICAL HISTORY is remarkable for chronic lung disease due to **smoking** .
SMOKER	11. history of **cigarette smoking** ,
SMOKER	He has a sixty to seventy five pack year **smoking** history and drinks alcohol approximately one time per week .
PAST-SMOKER	He is not a current **smoker** .
PAST-SMOKER	She quit **smoking** nine years ago .
PAST-SMOKER	The patient quit **tobacco** 45 years ago .
CURRENT-SMOKER	She **smokes** two to three packs per day times 30 years .
CURRENT-SMOKER	Please attempt to quit **smoking** .
CURRENT-SMOKER	**Smokes** one pack per day x 40 years .

Table 3: BM25 Similarity (sim_{BM25}).

$$sim_{BM25}(S^3_i, S^3_t) = \sum_{t \in T} (W_d \times W_q),$$

where,

$$W_d = \frac{(k_1+1)tf}{k_1((1-b) + b \times dl/avdl)},$$

$$W_q = \log \frac{N - n + 0.5}{n + 0.5}.$$

In this formula, T is the set of words appearing in the both S^3s, tf is the number of occurrences of a word t, dl is the length of S^3_t, $avdl$ is the average length of the S^3_t, N is the total number of S^3_t, n is the number of extracted S^3_t, and k_1 and b are the constants determined from the preliminary experiments. (We used $k_1 = 1.5$ and $b = 0.75$).

Table 4: The Number of Smoking Status.

Status	# of Status
UNKNOWN	252
SMOKER	9
CURRENT SMOKER	35
NON SMOKER	66
PAST SMOKER	36

1. **BASELINE1**: a majority-baseline. If the system could extract S^3 from a record, the system outputs NON-SMOKER, which is the most popular class among S,C,N and P. Otherwise, the system outputs UNKNOWN.
2. **BASELINE2**: this method uses character-based edit distance similarity (does not use BM25 similarity).
3. **PROPOSED**: the proposed method with various k values.

Result

The results are shown in Table 5. As shown in the table, the proposed system ($k = 10$) achieved the highest score, demonstrating the basic feasibility of the proposed approach.

Error Analysis

Table 6 shows error examples. Some errors come from rare expressions (singletons), such as "*off-*

Table 5: Results

Methods	Accuracy
BASELINE1	77.94%
BASELINE2	86.02%
PROPOSED ($k = 1$)	81.61%
PROPOSED ($k = 3$)	82.35%
PROPOSED ($k = 5$)	87.50%
PROPOSED ($k = 10$)	**88.97%**
PROPOSED ($k = 15$)	88.23%
PROPOSED ($k = 20$)	86.76%

and-on", which appears only once in the corpus. The system is poorly applicable to such rare words. Although the system handles such rare words poorly, a simple way to cope with this problem is a larger training set.

Other errors come from long S^3s, such as the following example:

> "*The patient is an 82 year-old right handed gentleman who has a past medical history of hypertension and **tobacco** use presented to the emergency room with acute change in mental status* ".

This S^3 includes several words that are not related to smoking status. Such unrelated words have a detrimental effect on the BM25 similarity. To cope with this problem, we must employ a more precise information extraction method, which captures only important expressions, in the near future.

Conclusion

The present paper introduced the proposed system, which classifies the smoking status of a patient using the medical record of the patient. The system consists of two modules: (1) a heuristic-based information extraction module and (2) an Okapi-BM25 and K-Nearest Neighbor-based (KNN-based) classifier module. In experiments, we achieved 88.97% accuracy, demonstrating the basic feasibility of the proposed approach. To achieve higher accuracy, a new approach that can extract more precise smoking information is highly desired.

Acknowledgments

Part of this research is supported by Grant-in-Aid for Scientific Research of Japan Society for the Promotion of Science （Project Number:16200039, F.Y.2004-2007 and 18700133, F.Y.2006-2007） and the Research Collaboration Project with Japan Anatomy Laboratory Co.Ltd.

References

[1] Robertson SE, Sparck-Jones K. Relevance weighting of search terms. Journal of the American Society for Information Science. 1976;27(1):129–146.

[2] Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M. Okapi at TREC-3. In: Proceedings of the 3rd Text Retrieval Conference; 1995. p. 109–126.

[3] Cover T, Hart P. Nearest Neighbour Pattern Classification. 1967;IT13(1):1–27.

[4] Fukunaga K. Introduction to Statistical Pattern Recognition, Academic Press Inc; 1972.

[5] Yu K, Guan G, Zhou M. Resume Information Extraction with Cascaded Hybrid Model. In: Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL2005); 2005. p. 499–506.

[6] Freitag D, McCallum A. Information extraction with HMMs and shrinkage. In: Proceedings of AAAI99 Workshop on Machine Learning for Information Extraction; 1999. p. 31–36.

[7] Sitter AD, Daelemans W. Information extraction via double classification. Proceedings of ATEM03; 2003.

[8] Finn A, Kushmerick N. Multi-level boundary classification for information extraction. Proceedings of ECML04; 2004.

[9] Borkar V, Deshmukh K, Sarawagi S. Automatic segmentation of text into structured records. In: Proceedings of ACM SIGMOD Conference; 2001. p. 175–186.

[10] Kushmerick N, Johnston E, McGuinness S. Information extraction by text classification. IJCAI01 Workshop on Adaptive Text Extraction and Mining; 2001.

[11] Fujii A, Iwayama M, Kando N. Overview of Patent Retrieval Task at NTCIR-4. In: Proceedings of the fourth NTCIR 4 workshop; 2004. p. 225–232.

[12] Murata M, Kanamaru T, Shirado T, Isahara H. Using the K Nearest Neighbor Method and BM25 in the Patent Document Categorization Subtask at NTCIR-5. In: Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies; 2005. p. 324–332.

Table 6: Error Examples.

System Output	Smoking Status (Gold Standard)	Smoking Status Sentence (S^3)
NON-SMOKER	CURRENT-SMOKER	**Smoking** :
NON-SMOKER	CURRENT-SMOKER	Positive **smoking** history .
CURRENT-SMOKER	PAST-SMOKER	The patient was a prior off-and-on **smoker** but has quit in 01/19 .
PAST-SMOKER	CURRENT-SMOKER	Abnormal Pap Test , history of ; Anemia ; Arrhythmia ; Gastrointestinal Problem , history of ; Herpes Simplex , Non Vulvovaginitis , history of ; Infertility ; Maternal Obesity ; Stopped **Smoking** This Pregnancy , history of ; Thyroid Nodule ; Urinary Tract Infection
PAST-SMOKER	CURRENT-SMOKER	He admits to an approximately 25-50 pack year **smoking** history , and social alcohol use .
PAST-SMOKER	SMOKER	The patient is an 82 year-old right handed gentleman who has a past medical history of hypertension and **tobacco** use presented to the emergency room with acute change in mental status .
NON-SMOKER	SMOKER	history of **cigarette** use , post menopausal , hypercholesterolemia .