An Trial Report to NTCIR10 MedNLP -- extracting medical diagnostic term by machine learning

Kota Kanno, Kazuyoshi Osanai, and Kyoji Umemura
Toyohashi University of Technology
Department of Computer Science and Engineering
kanno@ss.cs.tut.ac.jp, osanai@ss.cs.tut.ac.jp, umemura@tut.jp

ABSTRACT

This paper explains our approach toward NTCIR10-MEDNLP[1] tasks and what kind of problem we have encountered. We have select term extraction tasks since we have some experience about keyword extraction[2]. Since it is hard to build accurate dictionary or lexicon for medical term, we aimed to use machine learning and large amount of roughly tagged medical corpus as learning data. However, we are unable to prepare the learning data, and thus unable to make our system work.

Team Name

ulab

Subtasks

Complaint and diagnosis task

Keywords

Machine learning, Statistical Testing

1. INTRODUCTION

If we could know the medical terms in medical documents without any additional information, these terms would be useful for document processing, such as retrieval or classification. Though he lexicon is sufficient information to tag the various kinds of terms, it is very expensive to prepare the lexicon of them. If we have large amount of tagged documents, which shows which term are medical documents, computers can learn the concept and will be able to tag the unseen documents. In NTCIR10-MEDNLP, we have very small example of medical terms and thus, we need some effort to make computer work. We have tried to generate large amount of tagged document, which can be used as learning data of machine learning, and use machine learning algorithm to accomplish the task. Tough we cannot generate ideal learning data, we will be able to generate learning data, which may have many missing term but may contain sufficient information for context of the terms. However, we are unable to prepare the learning data, and thus unable to make our system work.

2. OVERVIEW OF OUR APPROACH

The structure of our design is illustrated in Fig. 1. We are trying to use machine learning for the annotation. The features that we select are set of terms, one or more characters surrounding the terms and its frequencies in the name position (positive frequency) and not in name position (negative frequency). An example is shown in figure 2. This example is an item to judge whether "認知症" is the name to extract. Basic idea is that machine can estimate whether a string is name or not by the positive frequency and negative frequency. The learning algorithm that we select is a kind of online learning, which is capable to handle large amount of learning data. Since the character unigram or bigram are used as features, the required learning data should be large. All possible surrounding character unigram or bigram should be appeared in the learning data.

Since only insufficient amount of samples for this approach are available, we need to prepare them by other means. We have gathered the documents of newspaper[3] which will contains the name of diseases. It is fairly easy to gather this kind of documents since the newspaper contains the report of death or disease of famous or influential person. We have gathered the documents which has clear term of death or disease, such as「死去=」「告別式」「党葬」「銀行葬」for death, and「入院」「中毒」「急死」「感染」for desease.

Unlike medical documents, the newspaper has standard way to report. Therefore, it is reasonable amount of effort to write a program which detect the name of disease by rules.
Thus we thought we would be able to generate a reasonable learning data for this task.

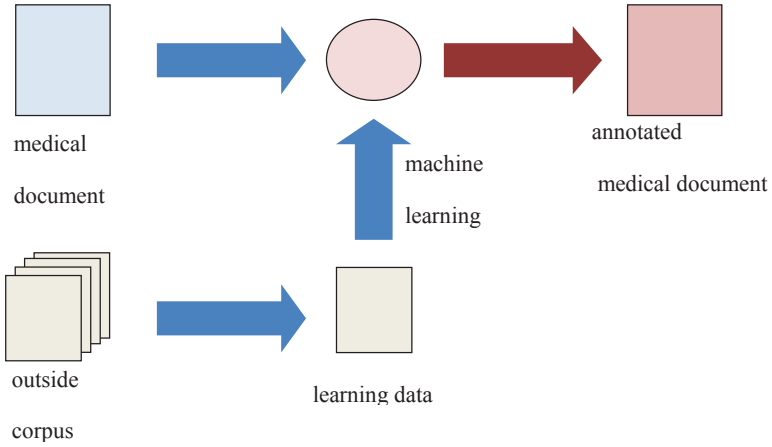

figure 1. system outline

3. WHAT HAPPENNED

After comparing the documents from newspapers and medical documents of NTCIR10, we have found that the term for one object is quite different between the two documents. Newspaper documents are aimed to be read by ordinal people, but medical documents are aimed to be read by professionals. For example a name of disease in medical documents is「ＭＰＯ－ＡＮＣＡ関連血管炎」, while this term will never appear in news paper documents since ordinal people never understand this.

Even worse, we assuming that the medical documents would have story like newspaper and this surrounding text may similar. For example we thought that the description like「突然の xx によって死亡した」would be common for newspaper document and medical documents. In fact, medical documents are usually not complete sentences since doctors are very busy and have only a little time to write the documents.

判定対象 よる、<c>認知症</c>が考えられた。
内部文字素性：（認知, 知症, 認, 知, 症）
　加えて，名前で内部で出現した頻度
　　名前外部で出現した頻度
外部文字組成：（よる，が考）
　加えて，名前の前後で出現した頻度
　　それ以外に出現した頻度

figuere 2. Exampe of features

4. SUBMITTED SYSTEM

After we have found that our system does not work properly, we try to make very simple system base on the following rules.

1. Record all terms in sample medical data.
2. Put modality "negation" to detect negative expression right after the term.
3. Put modality "suspicion" to find the character「疑」near the term.
4. Put modality "family" to find word showing family such as 「父、母」before the term.

The following is the sample of output of submitted system. Since it is very simple system but it may contribute as some bottom line.

This example of the result is shown in figure 3. Though we are not satisfied with this result, we choose not to resign the task..

・頭部ＭＲＩにて慢性虚血性変化あり、脳血管障害による、<c>認知症</c>が考えられた。
・筋性防御なし、
・φ１ｃｍ以下の<c>*狭窄*</c>部が３ｃｍ程あり。
・良性であっても<c>*巨大潰瘍*</c>の顕微鏡的悪性<c>*腫瘍*</c>可能性は
・②気管支<c>拡張</c>症

figuere 3. Exampe of submitted result

5. CONCLUSION

Though machine learning is powerful and generic tools, machine learning needs large amount of data in order to make system work. Though borrowing some untagged corpus of same genre may be possible, using difference genre of corpus is challenging and difficult approach. This report describes a trial to borrow the untagged corpus of different genre and report what kind of difficult problem exits for this particular task.

6. REFERECES

[1] NTCIR10 MedNLP(JP), http://www.mednlp.jp/medistj-ja/.
[2] Kazuyoshi Osanai and et al, Keyword Extraction and Linking from Newspaper Articles, C6-3, *Proceedings of Forum on Data Engineering and Information Management*(Fukuyama, Japan, March 03 – 05, 2013). DEIM2013. IEICE.
[3] Mainichi Shinbun 91,92,93,94,95,96,97