Mapping Clinical Notes to Medical Terminology at Point of Care

Yefeng Wang
School of Information Technologies
University of Sydney
New South Wales 2006, Australia
ywang1@it.usyd.edu.au

Jon Patrick
School of Information Technologies
University of Sydney
New South Wales 2006, Australia
jonpat@it.usyd.edu.au

Abstract

Clinicians write the reports in natural language which contains a large amount of informal medical term. Automating conversion of text into clinical terminologies allows reliable retrieval and analysis of the clinical notes. We have created an algorithm that maps medical expressions in clinical notes into a medical terminology. This algorithm indexes medical terms into an augmented lexicon. It performs lexical searches in text and finds the longest possible matches in the target terminology, SNOMED CT. The mapping system was run on a collection of 470,000 clinical notes from an Intensive Care Service (ICS). The evaluation on a small part of the corpus shows the precision is 70.4%.

1 Introduction

A substantial amount of clinical data is locked away in a non-standardised form of clinical language which if standardised could be usefully mined to gain greater understanding of patient care and the progression of diseases. Clinical notes on a patient's health are written in natural language which contains a great deal of formal terminology but used in an informal and unorderly manner. These medical notes need to be converted to a formal terminology to enable accurate retrieval and to compile aggregated statistics of the medical care. To satisfy these needs, we developed a medical concept identifier that is able to identify concepts in clinical notes and mapped to medical codes in a terminology. The algorithm has been implemented to tag medical concepts in a collection of 470,000 clinical notes from an Intensive Care Service. A total of 9,135,000 instances of about 20,000 medical concepts were identified. These medical concepts are used to study the medical language used by Intensive Care clinical staff, and the identified concepts are used to index patient clinical records for targeted information retrieval activities.

2 Related Work

There has been a large effort spent on automatic recognition of medical and biomedical concepts and mapping them to medical terminology. The Unified Medical Language System Meta-thesaurus (UMLS) is the world's largest medical knowledge source and it has been the focus of much research. One of the prominent systems to map free text to UMLS are MetaMap (Aronson, 2001),

3 Constructing the Lexicon

The Augmented Lexicon is a data structure developed to keep track of the words that appear in the concepts of the medical terminology. The Augmented Lexicon is built from the individual words in the gloss or the definition of the medical term. For example, *Myocardial Infarction* has the atomic words *Myocardial* and *Infarction*. Each concept is normalised which includes removal of stop words, stemming, and spelling variation generation. For each word, a list of the concept ids that contain that word is stored in the Augmented Lexicon. An additional table is stored alongside the augmented lexicon, called the "atomic term count" to record the number of atomic terms that comprise each description.

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 102–103,
Columbus, Ohio, USA, June 2008.

4 Token Matching Algorithm

The algorithm performs string alignment between the source text and a target medical terminology. The best matches are determined by scoring algorithms for both perfect matching and partial matching. To find all possible matches, the algorithm iteratively performs matches for sub-strings using dynamic programming, so that the algorithm doesn't have to generate all combination of sub-strings for the input sentence. Each previously computed substrings matches are stored and in a matching matrix so don't require recalculation.

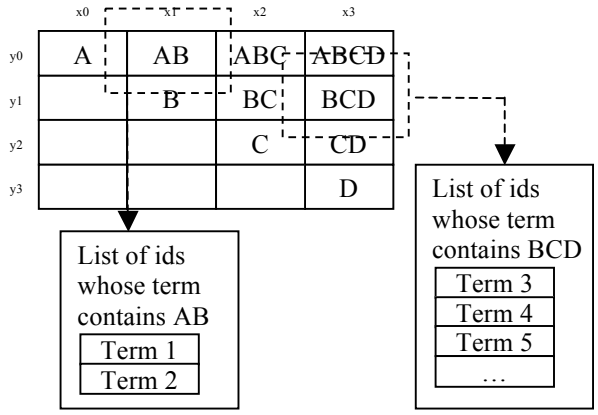

Figure 1: Token Matching Matrix

The data stored in each cell is a list of medical term ids that are in all the tokens that comprise the cell. The score is then calculated using the "atomic term count", which stores the number of tokens that make up that term. The score is the number of tokens in the current cell that have the term id in common divided by the number of tokens in the full description.

5 Recognition of Clinical Entities

Before medical term identification, Clinical entities such as measurement, demography, quantities are recognised and normalised to their classes.

Entity Class	Examples
Blood Pressure	105mm of Hg
Demography	69 year-old man
Datetime	20/11 2030
Quantity	55 mm

Table 1. Clinical Entities and Examples

6 Evaluation

The token matching algorithm has been implemented as a module in a terminology server that can provide real time text to medical concept encoding. The system was installed in the Intensive Care Service that provides web interfaces for users to submit clinical notes and it computed SNOMED CT codes in real-time. The web interface has been implemented in several clinical forms templates at the RPAH, allowing data to be captured as the doctors fill in these forms. A feedback form has been implemented allowing clinicians to submit comments, identify terms that are missed by the system and submit corrections to incorrectly labeled terms. This was seen as a rare opportunity to collect an expert corrected corpus of clinical notes. Unfortunately, there was little adherence to the correction part of the program and so we do not yet have sufficient material to be precise about recall values.

To evaluate the accuracy our systems, we collected a set of bedside clinical notes of patient monitoring chart information. 487 documents and 4,054 medical concepts were tagged with SNOMED CT codes and have been evaluated by medical experts. There are 2,852 correctly identified concepts and 1,202 incorrectly identified concepts, results in a precision rate of 70.4%. The recall rate hasn't been fully evaluated.

7 Conclusions

In conclusion, we have proposed a system to find medical terms in free text clinical notes and map them into a medical terminology. We have implemented the algorithm as a web-service system. The algorithm uses an augmented lexicon to index concept descriptors in SNOMED CT, which allow a much faster mapping of longest spanning concepts in the system than a naïve word searching approach, which can then create more effective information retrieval and information extraction.

References

Aronson, A. R. (2001). *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program*. Proc AMIA Symp 17: 21.