Extracting Patient Clinical Profiles from Case Reports

Yitao Zhang and **Jon Patrick**
School of Information Technology
University of Sydney
NSW 2006, Australia
{yitao, jonpat}@it.usyd.edu.au

Abstract

This research aims to extract detailed clinical profiles, such as signs and symptoms, and important laboratory test results of the patient from descriptions of the diagnostic and treatment procedures in journal articles. This paper proposes a novel mark-up tag set to cover a wide variety of semantics in the description of clinical case studies in the clinical literature. A manually annotated corpus which consists of 75 clinical reports with 5,117 sentences has been created and a sentence classification system is reported as the preliminary attempt to exploit the fast growing online repositories of clinical case reports.

1 Corpus and Mark-up Tags

This paper proposes a mark-up scheme aimed at recovering key semantics of clinical case reports in journal articles. The development of this mark-up tag set is the result of analysing information needs of clinicians for building a better health information system. During the development of this tag set, domain experts were constantly consulted for their input and advice.

1.1 The Mark-up Tag Set

- **Sign** is a signal that indicates the existence or nonexistence of a disease as observed by clinicians during the diagnostic and treatment procedure. Typical signs of a patient include the appearance of the patient, readings or analytical results of laboratory tests, or responses to a medical treatment.
- **Symptom** is also an indication of disorder or disease but is noted by patients rather than by clinicians. For instance, a patient can experience weakness, fatigue, or pain during the illness.
- **Medical test** is a specific type of sign in which a quantifiable or specific value has been identified by a medical testing procedure, such as blood pressure or white blood cell count.
- **Diagnostic test** gives analytical results for diagnosis purposes as observed by clinicians in a medical testing procedure. It differs from a medical test in that it generally returns no quantifiable value or reading as its result. The expertise of clinicians is required to read and analyse the result of a diagnostic test, such as interpreting an X-ray image.
- **Diagnosis** identifies conditions that are diagnosed by clinicians.
- **Treatment** is the therapy or medication that patients received.
- **Referral** specifies another unit or department to which patients are referred for further examination or treatment.
- **Patient health profile** identifies characteristics of patient health histories, including social behaviors.
- **Patient demographics** outlines the details and backgrounds of a patient.
- **Causation** is a speculation about the cause of a particular abnormal condition, circumstance or case.
- **Exceptionality** states the importance and merits of the reported case.

Total articles	75
Total sentences	5,117
Total sentences with tag	2,319
Total tokens	112,382
Total tokens with tag	48,394

Table 1: Statistics of the Corpus

- **Case recommendations** marks the advice for clinicians or other readers of the report.

- **Exclusion** rules out a particular causation or phenomenon in a report.

1.2 The Corpus

The corpus described in this paper is a collection of recent research articles that report clinical findings by medical researchers. To make the data representative of the clinical domain, a wide variety of topics have been covered in the corpus, such as cancers, gene-related diseases, viral and bacteria infections, and sports injuries. The articles were randomly selected and downloaded from BioMed Central[1]. During the selection stage, those reports that describe a group of patients are removed. As a result, this corpus is confined to clinical reports on individual patients. A single human annotator (first author) has manually tagged all the articles in the corpus. The statistical profile of the corpus is shown in Table 1.

2 The Sentence Classification Task

The patient case studies corpus provides a promising source for automatically extracting knowledge from clinical records. As a preliminary experiment, an information extraction task has been conducted to assign each sentence in the corpus with appropriate tags. Among the total of 2,319 sentences that have tags, there are 544 (23.5%) sentences assigned more than one tag. This overlapping feature of the tag assignment makes a single multi-class classifier approach not appropriate for the task. Instead, each tag has been given a separate machine-learned classifier capable of assigning a binary 'Yes' or 'No' label for a sentence according to whether or not the sentence includes the targeted information as defined by the tag set. Meanwhile, a supervised-learning approach was adopted in this experiment.

[1]http://www.biomedcentral.com/

Tag	Precision	Recall	F_1
Diagnostic test	66.6	46.8	55.0
Medical test	80.4	51.6	62.9
Treatment	67.6	44.7	53.8
Diagnosis	62.5	33.8	43.8
Sign	61.2	50.5	55.4
Symptom	67.8	45.8	54.7
Patient demographics	91.6	73.1	81.3
Patient health profile	53.0	24.1	33.2

Table 2: Sentence Classification Result for Some Semantic Tags

A Maximum Entropy (MaxEnt) classifier[2] and a SVM classifier (SVM-light) with tree kernel (Moschitti, 2004; Joachims, 1999) were used in the experiment. The SVM classifier used two different kernels in the experiment: a linear kernel (SVM t=1), and a combination of sub-tree kernel and linear kernel (SVM t=6). The introduction of the tree kernel was an attempt to evaluate the effectiveness of incorporating syntactic clues for the task. The feature set used in the experiment consists of unigrams, bigrams, and title of the current section. The experiment results for selected markup tags are shown in Table 2.

Acknowledgments

We wish to thank Prof Deborah Saltman for defining the tag categories and Joel Nothman for refining their use on texts.

References

Thorsten Joachims. 1999. Making Large-scale SVM Learning Practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods-Support Vector Learning*.

Alessandro Moschitti. 2004. A Study on Convolution Kernels for Shallow Semantic Parsing. In *Proceedings of the 42-th Conference on Association for Computational Linguistic*.

[2]Zhang Le http://homepages.inf.ed.ac.uk/s0450736/