Extracting Semantics in a Clinical Scenario

Yitao Zhang **Jon Patrick**

School of Information Technologies
University of Sydney
NSW 2006, Australia
Email: yitao, jonpat@it.usyd.edu.au

Abstract

Unlike abstracts, full articles of clinical case studies provide more detailed profiles of a patient, such as signs and symptoms, and important laboratory test results of the patient from the diagnostic and treatment procedures. 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.

Keywords: Clinical Profile, Semantics, Natural Language Processing

1 Introduction

Two fundamental questions dominate daily practice of clinicians: first of all, given the history of medical tests and symptoms of a patient, what is the best possible explanation or diagnosis of the condition? Secondly, what is the best treatment under the specified circumstances? Both questions are generally answered with individual expertise of medical practitioners. However, in environments like an intensive care unit of a hospital, where there is 24-hour monitoring of patients who are in critical conditions and real-time interpretation of laboratory tests are needed, an information system which can automatically retrieve the most similar cases to the current patient from existing databases or knowledge bases will be of great value for helping clinicians to make the right decisions in a timely manner.

Following the paradigm of evidence-based medicine (EBM), which uses best available research information to support clinical decision making, people try to learn domain-specific knowledge from three different types of clinical texts, namely patient records, case studies, and clinical aggregation and summarisation of a collection of like cases.

- **Patient Records** store raw clinical notes which are generated at different stages of the diagnostic and treatment procedure of a patient. Although it can provide first-hand the richest information about patients, their confidentiality has always been a critical problem for the research community to have access to enough data for developing sophisticated learning systems comparable to human performance.
- **Case Studies** provide detailed discussions of patients with abnormal symptoms or interesting conditions which are considered as report-worthy by medical researchers. Most case studies are published in scientific journals as research papers and are often publicly available through online repositories like BioMed Central and MEDLINE.
- **Clinical Summarisation** provides highly accurate and relevant clinical knowledge compiled by human experts by reviewing and summarising the latest progress in the field. However, the cost of maintaining such a knowledge base is very high due to the fact that human experts are not capable of reviewing every single case report. Moreover, manually compiled summarisations are generally unable to provide correct answers when extra conditions are added into a live case. For instance, a query on the treatment of a common disease is prone to fail if it is accompanied by extra signs, symptoms, or disease conditions, as noted by Ely et.al. (2000, 2005) in their research to develop a taxonomy for clinical questions.

Until now, most efforts have been made to extract clinical knowledge from only journal abstracts, which is still a small proportion of the full text available. Important clinical information, such as laboratory tests and readings, signs and symptoms of a patient during the diagnosis and treatment, and health profile of a patient, is generally missing in the abstract of a case report. In order to overcome this problem, this paper turns to full research articles of clinical reports for detailed descriptions of diagnoses and treatments by medical practitioners. The ultimate goal of this research is to develop an information retrieval system which is aware of the uniqueness of each individual patient case in performing retrieval on full journal articles of patient case studies. As the first step of the study, this paper will focus on using Natural Language Processing (NLP) techniques to extract targeted information of patient clinical profiles from text, which provide crucial clues for differentiating individual patients. The system reported in the paper is expected to provide an intermediate level of domain specific abstraction of information, and therefore to fill a gap between the need for deeper understanding of the text, and the difficulty of coping with a large amount of data in a clinical scenario.

This paper will first discuss the process of creating a new patient case report corpus from online journal articles. These articles report interesting patient cases and contain detailed descriptions of the diagnostic and treatment procedure of each patient which provides a rich source for extracting clinical profiles.

A novel mark-up tag set has been proposed to cover a wide variety of semantics in the descriptions of diagnostic and treatment procedures by clinicians. Possible applications of the proposed corpus include Information Extraction, Information Retrieval, and Question Answering. As a preliminary attempt, a sentence classification task has been set up to train a system that can automatically assign semantic information to text.

2 Related Works

Applying natural language processing techniques to the clinical domain has recently attracted much attention. In the Clinical e-Science Framework (CLEF) project, the chronicle of conditions and treatments of a patient was created by combining various information segments from a repository of unstructured patient records, such as admission summaries, referral letters, and radiology and pathology reports. (Harkema, Roberts, Gaizauskas & Hepple 2005) In this work the recognition of temporal expressions and correct linking with their corresponding events in medical narratives played a key role in the integration of extracted information into a single diagnostic and treatment history of a patient.

In an attempt to develop a clinical question-answering system, Niu and Hirst (2004) tried to extract knowledge of clinical outcomes from Clinical Evidence (CE), which is a constantly updated publication of clinical findings. Demner-Fushman and Lin (2006*a*, 2006*b*) used CE as the gold-standard source for evaluating the performance of their system on answering clinical questions like "What is the best drug treatment for X?" from abstracts in MEDLINE.

To generate better text summarisations, sentence classification methods have been used to first reveal the rhetorical structure of scientific articles, which encodes the argument role of each sentence in a text, such as "introduction", "method", "result", and "conclusion". (McKnight & Srinivasan 2003) Discriminative methods like Support Vector Machines (SVM) have been widely used in the sentence classification task in the biomedical domain. (McKnight & Srinivasan 2003, Yamamoto & Takagi 2005) Lin et.al. (2006) recently tried to model the rhetorical structure of MEDLINE abstracts by using a generative model.

Hara and Matsumoto (2005) used BACT, which is a machine learner that can capture patterns from semi-structured data like parse trees, to classify sentences in MEDLINE abstracts as to whether or not a sentence contains the information extraction targets like "compared treatment", "endpoint", and "patient population". The BACT system is reported as comparable to a SVM learner with tree kernel. (Kudo & Matsumoto 2004)

3 Corpus and Mark-up Tags

3.1 The Mark-up Tag Set

This section proposes a mark-up tag set aimed at recovering key semantics of clinical case reports in journal articles. The development of this mark-up scheme is the result of analyzing 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.

The mark-up tags can be further categorized into two sub-groups, namely the more generic tags representing descriptions of the diagnostic and treatment procedures of clinicians, and the genre-specific tags reflecting the argumentation role of medical research papers. Examples of the tagged sentences are shown in Table1 and Table2.

Sign	"his finger was swollen, tense and tender."
	"electrolytes were normal"
	"Therefore, she did not respond to neither topical nor systemic steroid."
Symptom	"she also complained of right upper quadrant pain, which radiated to her right lower quadrant and upper back."
	"he experienced mild muscle weakness in his lower extremities"
	"At the age of 33 years he again presented with extreme fatigue"
Medical test	"Her BP was 100/75 mmHg, HR was 110 bpm"
	"respiratory rate was 20 breaths / min"
Diagnostic test	"Cerebrospinal fluid examination was negative"
	"CT scan with contrast confirmed the vascular nature of the cyst."
Diagnosis	"The diagnosis of Echinococcus granulosus infection was confirmed"
	"This report describes a case of cholestatic jaundice"
Treatment	"He was placed on triple antibiotics therapy."

Table 1: Marked-up Text for Diagnostic and Treatment Procedure

3.1.1 The Diagnostic and Treatment Procedure

The mark-up tags described in this section try to convey a variety of general concepts which occur in descriptions of clinical procedures. These tags are more generic and are not confined to any specific genre of medical text, such as research articles or raw clinical notes.

- **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. Clinical tests which are of particular interest to medical practitioners were given separate mark-up tags in this research.
- **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. In most cases, descriptions of symptoms are expressed as complaints or feelings by patients. However, an interesting exception can be seen in the following sentence:

  > "He was identified by nursing staff as having knee pain."

  In this case a patient with dementia actually lost his ability to feel pain, and his symptom was only noted by clinicians through their observations.

- **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.

3.1.2 Genre Specific Semantics

The mark-up tags described in this section provide genre-specific semantics. These tags tend to occur much more frequently in journal articles of clinical findings than other types of medical text like clinical notes.

- **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.
- **Case recommendations** marks the advice for clinicians or other readers of the report.
- **Exclusion** rules out a particular causation or phenomenon in a report.

3.2 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], which is a freely available online repository of biomedical research papers. 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. Annotations were done by using Callisto, which is a free annotation tool developed by MITRE.[2] The statistical profile of the corpus, together with the distribution of mark-up tags, are shown in Table 3 and Table 4, respectively. The mark-up tag set covers 45.3% of all the sentences in the corpus. The sentences that cannot be properly classified are not assigned with any semantic tag. Manual inspection of the corpus shows that most sentences not covered by the mark-up tag set are those that describe background knowledge of the research, or reference the work of other researchers.

[1] http://www.biomedcentral.com/
[2] http://callisto.mitre.org/

Referral	"The patient was referred to the pediatric surgery department"
Patient health profile	"this healthy , well-nourished , normotensive postmenopausal woman"
	"She had a history of symptomatic therapy (non-specific antibiotics)."
Patient demographics	"A 49 year old woman"
Causation	"due to a malassezia furfur infection on the skin"
	"The rash was thought to be a reaction to the Cephalosporin antibiotic."
Exceptionality	"Our case is the first male patient reported with SSc and IPH. "
	"Situs inversus presenting with acute cholecystitis is very rare."
Case recommendations	"Conservative treatment should be given unless there are complications."
	"Glucocorticoids should be avoided in diabetic patients."
Exclusion	"She has no history of smoking."
	"There was no current or history of asthma."

Table 2: Marked-up Text for Genre Specific Semantics

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

Table 3: Statistics of the Corpus

3.3 Inter-annotator Agreements

Kappa statistics are the most commonly used measure for evaluating inter-annotator agreements. The κ value is defined as $\kappa = \frac{P(A)-P(E)}{1-P(E)}$, in which $P(A)$ is the observed annotator agreement, and $P(E)$ is the expected agreement by chance. The upper bound of the κ value is 1, which means perfect agreement, and the lower bound -1 suggests totally disagreement. In evaluation of the quality of a linguistic corpus, a κ value of 0.67 is generally considered as a substantially high level of agreement.

Due to the difficulty of assembling more human annotators, there were only two people (including the first author) involved in the experiment of evaluating the inter-annotator agreement. A small proportion of the corpus (5 documents with 302 sentences and 7,150 tokens) has been randomly selected and tagged by both annotators. The κ values are calculated on both the word and the sentence level. For the sentence level, agreement was recorded if both annotators agree that the current sentence contains targeted information as specified by a certain mark-up tag. For the word level agreement, the two annotators have to agree on the exact scope of the selected words for each

Tag	Total Sentences	Percentage
Sign	955	18.7%
Treatment	460	9.0%
Diagnostic test	416	8.1%
Diagnosis	217	4.2%
Medical test	176	3.4%
Case recommendations	171	3.3%
Patient health profile	166	3.2%
Symptom	145	2.8%
Patient demographics	141	2.8%
Exceptionality	72	1.4%
Causation	64	1.3%
Exclusion	29	0.6%
Referral	22	0.4%

Table 4: Distribution of Mark-up Tags

Tag	Sentence kappa	Word kappa
Referral	1	0.3
Patient demographics	0.94	0.91
Diagnosis	0.76	0.45
Symptom	0.75	0.61
Case recommendations	0.63	0.61
Treatment	0.61	0.54
Case category	0.6	0.4
Sign	0.58	0.57
Causation	0.57	0.49
Patient health profile	0.56	0.55
Diagnostic test	0.54	0.4
Medical test	0.47	0.4
Exceptionality	0.44	0.5
Case circumstances	0.41	0.31
Exclusion	0.15	0.02

Table 5: Inter-annotator Agreement

mark-up tag. For example, if the first annotator labeled "swollen" in "his finger was swollen, tense and tender." as a Sign, while the second annotator only selected "tense and tender" as a Sign, then they have a perfect agreement on the sentence level while totally disagree with each other on the word level.

For each individual mark-up tag, a κ value is calculated on both the sentence and the word level. If the mark-up tag has been assigned to some words in a sentence, a 'Yes' label will be attached to those words and the sentence, otherwise a 'No' label is used. The assumption of different distributions among annotators (Cohen 1960) was applied in the calculation of $P(E)$, the expected agreement by chance. For example, for the mark-up tag 'Sign' in a corpus of the size of 10 sentences with 100 words, if the first annotator has selected 70 words which distribute in 8 unique sentences as positive instances of 'Sign', and the second annotator has labeled 60 words in 5 unique sentences, then their expected agreement by chance can be calculated as follows:

$P(E)_{sentence\ level} = \frac{8}{10} \times \frac{5}{10} + \frac{2}{10} \times \frac{5}{10} = 0.5$

$P(E)_{word\ level} = \frac{70}{100} \times \frac{60}{100} + \frac{30}{100} \times \frac{40}{100} = 0.54$

The detailed results of both sentence and word level agreement for all mark-up tags are shown in Table 5.

Although a high κ value does not necessary mean perfect annotation, it still shows the stability of the proposed annotation framework. In Table 5, "patient demographics" achieves above 0.9 on κ value at both sentence and word levels, which suggests a high level of annotator agreement. Although "referral" has a sentence level κ value of 1, it shows a poor agreement on the word level which indicates the two annotators may not be referring to the same fact while they both agree the same type of information about a "referral" occurs within the same sentence. Most other mark-up tags have κ values above 0.4 at both sentence and word levels which shows an intermediate level of annotator agreement. The "exclusion" tag has the lowest κ value, particularly on the word level, and therefore reveals the ambiguous nature of the tag.

4 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. This strategy has potential for further use as the knowledge base for building better text-mining systems for health management purposes.

Although the current mark-up tag set covers a good variety of aspects of the semantics of clinical reports, it does not provide a hierarchy or rhetorical structure of text. Instead, many sentences are considered as containing information that serves more than one functionality. For instance, the following sentence contains both information of the treatment "pneumonectomy", and the diagnosis "bronchogenic carcinoma".

> She was admitted for pneumonectomy with a provisional diagnosis of bronchogenic carcinoma.

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.

4.1 Methods

4.1.1 Preprocessing

Raw articles were first separated into different sections based on their original formatting information. A maximum entropy based sentence splitter was then used to detect the boundaries between sentences. Each sentence was part-of-speech (POS) tagged and chunked by using Genia Tagger (Tsuruoka, Tateishi, Kim, Ohta, McNaught, Ananiadou & Tsujii 2005) which is trained on biomedical text. The POS information was also used as the input for the Bikel parser (Bikel 2004) which is trained on newswire text. The system used AGTK[3], an API interface developed for accessing annotation graph compatible file formats, to retrieve human annotations made within Callisto. Figure 1 shows a snapshot of the GUI interface for human annotations.

[3] http://agtk.sourceforge.net/

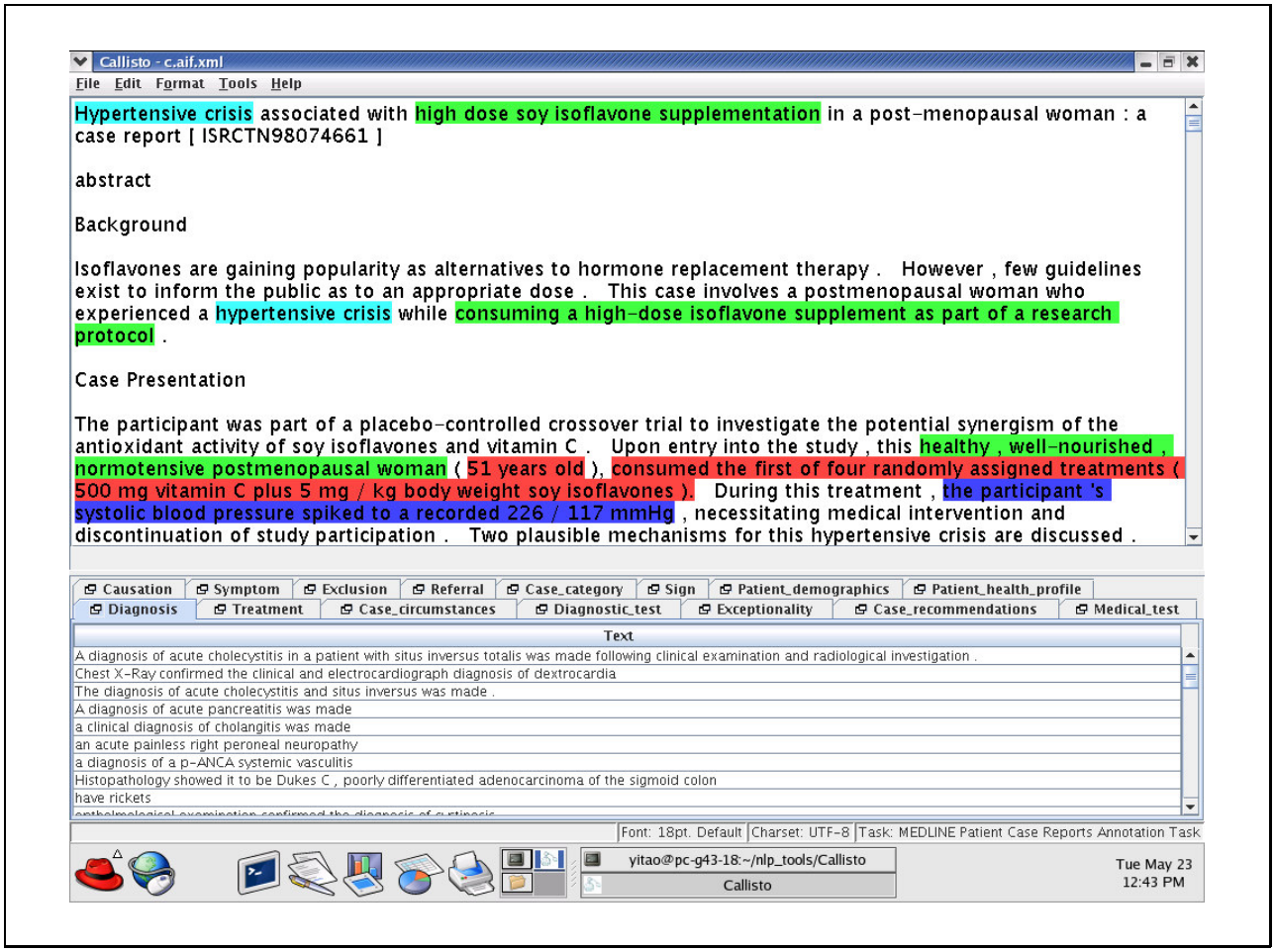

Figure 1: The Callisto GUI for Manual Annotations

4.1.2 Classifiers and the Feature Set

Maximum Entropy Modeling (MaxEnt) and Support Vector Machine (SVM) are the two most popular techniques for supervised classification problems. This section briefly discusses the principles of the two methods, together with the feature set and experimental settings.

In a MaxEnt model, features are considered as constraints. (Berger, Pietra & Pietra 1996) For example, when considering the mark-up tag "Sign", the system tries to assign either a "YES" or a "NO" label to a sentence based on various features of the instance, such as

$$f_i(y,x) = \begin{cases} 1 & \text{if } y\text{=YES and has\_word}(x)\text{='swollen'} \\ 0 & \text{otherwise} \end{cases}$$

where y is the outcome class label for the mark-up tag "Sign", and x is the current sentence instance. In this example, the feature fires when the current sentence contains the unigram word "swollen". The conditional probability of seeing the class label y given an instance x can be then defined in terms of its features:

$$P(y|x) = \frac{1}{Z(x)} exp\left(\sum_i \lambda_i f_i(y,x)\right) \qquad (1)$$

where $Z(x)$ is a normalization function, and each feature f_i is associated with a parameter λ_i. An important advantage of MaxEnt modeling is that it can tolerate dependent features within the same feature set. In the experiments, we used a MaxEnt classifier implementation which is developed by Zhang Le.[4].

[4]http://homepages.inf.ed.ac.uk/s0450736/

Given a set of instances $x_i \in R^n$, where R^n is a N dimensional vector space and each x_i is assigned with a binary class label of either -1 or +1, an SVM classifier tries to project input vectors onto a high-dimensional feature space Γ by using a mapping function called a kernel:

$\Phi : x_i \rightarrow \Gamma$

It then separates two classes in the high-dimensional feature space Γ by using a linear hyperplane with the maximum margin. (Vapnik 1995)

The tree kernel proposed by Collins and Duffy (2002) exploits non-flat structure representations and is able to measure the similarity between two parse trees by comparing all their common substructures. By combining the tree kernel with other conventional kernels such as linear, polynomial, and radial basis function (RBF) kernels, an SVM classifier is able to not only utilize common flat-structured features, but also exploit syntactic clues from parse trees without any explicit rules for defining syntactic features. In the experiments, a SVM-light package with tree kernel (Moschitti 2004, Joachims 1999) was used. The SVM classifier used two different kernels in the experiment: a linear kernel (SVM t=1), and a combination of the tree kernel and the linear kernel (SVM t=6). The introduction of the tree kernel was an attempt to evaluate the effectiveness of incorporating syntactic features for the task.

The feature set used in the experiment consists of bag-of-words features (unigrams and bigrams) and the title information of the current section. The syntactic features were encoded in parse trees and were therefore only accessed implicitly by the SVM classifier with the tree kernel.

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
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 6: Sentence Classification Result for Some Semantic Tags

Features	Classifier	Precision	Recall	F_1
	Baseline	18.7	100.0	31.5
Unigram	MaxEnt	56.8	48.2	52.1
	SVM t=1	53.6	36.6	43.5
	SVM t=6	73.5	34.0	46.5
Unigram +Bigram	MaxEnt	58.5	47.4	52.4
	SVM t=1	55.6	25.2	34.7
	SVM t=6	74.7	29.4	42.2
Unigram +Section Title	MaxEnt	59.0	50.1	54.2
	SVM t=1	61.6	46.8	53.2
	SVM t=6	70.5	40.9	51.7
Unigram +Bigram +Section Title	MaxEnt	61.2	50.5	55.4
	SVM t=1	61.7	39.3	48.0
	SVM t=6	71.3	36.1	47.9

Table 7: Sentence Classification Result for "Sign"

4.1.3 Evaluation Metrics

The evaluation metrics are standard $Precision = \frac{tp}{tp+fp}$, $Recall = \frac{tp}{tp+fn}$, and $F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$ in which tp is the number of true positive instances, fp is the number of false positives, and fn is that of false negative examples as decided by a classifier.

4.2 Experiment Results

In order to evaluate our system in more detail, this section will mainly focus on the "Sign" tag which has the largest occurrences in the corpus and is relatively stable according to the inter-annotator agreement test results. The sentence classification result for the "Sign" tag is shown in Table 7. The figures in the table are 10-fold cross validation results. The baseline system reported in the table uniformly predicts positive for all sentences.

The MaxEnt classifier generally outperforms the SVM classifier in every test on the F1 metric. However, the difference between two systems is not statistically different in some experimental settings based on 10-fold paired t-test results for the significance level of $\alpha = 5\%$. The introduction of the tree kernel in the SVM classifier shows a higher precision rate, which suggests the potential benefits of exploiting syntactic features in the task, and a slight drop in recall rate compared to the MaxEnt system and the SVM system with linear kernel alone. However, the tree kernel suffers the problem of inaccurate input parse trees because of the lack of gold-standard medical tree-banks on which to train a parser. The Bikel parser is trained on newswire text and therefore much less accurate on parsing clinical reports.

The recall rate of SVM classifiers drops when bigrams are added to unigram features, while the MaxEnt classifier shows no significant improvement. Although Niu et.al. (2005) reported better performance by incorporating bigram features, this sentence classification experiment showed no significant benefits by using bigrams.

Adding title information of the current section into the feature set has given both SVM and MaxEnt classifiers significant improvement in terms of precision and recall. It suggests that the rhetorical role of a sentence, revealed partly by its section title, plays an important role in differentiating the functionality of different text segments. Moreover, this feature can be easily acquired in full journal articles.

Sentence classification results for other mark-up tags are shown in Table 6. The figures shown in the table are MaxEnt classifier results using only section title and unigram features. Due to the data sparseness problem, only 4-fold cross validation is used in these experiments.

4.3 Discussion

The imbalanced class sizes of the targeted information in the corpus makes the sentence classification a difficult task. Most classification experiments have an F1 value well below 60. In this section we review some of the wrongly classified sentences and try to identify strategies for improving the current information extraction system.

First of all, simple bag-of-words modeling of text is not able to capture the rich and subtle information encoded in the sentences. For example, the sentence

> His mother and his brother had an history of deep venous thrombosis and his father died because of pulmonary embolism.

was misclassified as a positive instance of "Sign" by both classifiers. This sentence actually mentions the family health history of the patient and therefore includes the "health profile" information as decided by human annotators. However, a machine learner with only bag-of-words modeling of text failed to recognise the difference between the switching of person entities from patient to his family members.

Moreover, the manual annotation also introduces some inconsistencies and errors, which will need to be corrected for future annotation. The sentence

> Ultrasonographic examination of the mass showed a cystic structure.

is classified as "Sign" by the system. However, in manual annotations, since the sentence is already given a "Diagnostic test" tag, it is therefore not tagged as "Sign" by human annotator. Although most sentences with "Diagnostic test" or "Medical test" information include "Sign" information too, there are some examples where no description of any sign or symptom occurred with medical test. This problem can be avoided by introducing more rigid and consistent guidelines on manual annotation.

5 Conclusion and Future Work

The fast-growing content of online repositories like BioMed Central and MEDLINE provides a valuable source for learning sophisticated text-mining system for the clinical domain. This paper aims to evaluate the possibility of exploiting full journal case studies rather than only abstracts which generally ignore the detailed clinical history of each individual patient. By using simple bag-of-words features and rhetorical structure information, machine learned classifiers

can easily outperform a naive baseline which uniformly predicts true in the sentence classification task in which each sentence is assigned tags according to the information it contains about the diagnostic and treatment procedure.

Future work includes annotating more case reports by inviting domain experts to participate in the project. More annotators are needed for evaluating the stability of the mark-up tag set on a larger scale. We are also considering adding more specialized semantic tags like "outcomes" of a treatment into the current tag set. In order to achieve better performance on the sentence classification task, we plan to incorporate domain knowledge which is partly encoded in biomedical ontologies like Snomed CT and UMLS. Successful mapping between free text to its corresponding clinical concepts would be critical for the system. Moreover, syntactic features need to be evaluated and refined to provide more useful clues of the text.

Acknowledgments

We wish to thank Prof Deborah Saltman for defining the tag categories and Joel Nothman for refining their use on texts. We would also like to thank the support from all members of the Sydney Language Technology Group, particularly Dr James Curran and Yefeng Wang.

References

Berger, A. L., Pietra, S. D. & Pietra, V. J. D. (1996), 'A Maximum Entropy Approach to Natural Language Processing', *Computational Linguistics* **22**(1), 39–71.

Bikel, D. M. (2004), A Distributional Analysis of a Lexicalized Statistical Parsing Model, *in* 'EMNLP2004'.

Cohen, J. (1960), A Coefficient of Agreement for Nominal Scales, *in* 'Educational and Psychological Measurement', number 20, pp. 37–46.

Collins, M. & Duffy, N. (2002), New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete Structures, and the Voted Perceptron, *in* 'ACL2002'.

Demner-Fushman, D. & Lin, J. (2006*a*), Answer Extraction, Semantic Clustering, and Extractive Summarization for Clinical Question Answering, *in* 'Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Sydney, Australia, pp. 841–848.

Demner-Fushman, D. & Lin, J. (2006*b*), Situated Question Answering in the Clinical Domain: Selecting the Best Drug Treatment for Diseases, *in* 'Proceedings of the Workshop on Task-Focused Summarization and Question Answering', Association for Computational Linguistics, Sydney, Australia, pp. 24–31.

Ely, J. W., Osheroff, J. A., Chambliss, M. L., Ebell, M. H. & Rosenbaum, M. E. (2005), 'Answering Physicians' Clinical Questions: Obstacles and Potential Solutions', *Journal of the American Medical Informatics Association* **12**(2), 217–224.

Ely, J. W., Osheroff, J. A., Gorman, P. N., Ebell, M. H., Chambliss, M. L., Pifer, E. A. & Stavri, P. Z. (2000), A Taxonomy of Generic Clinical Questions: Classification Study, *in* 'British Medical Journal', 321, pp. 429–432.

Hara, K. & Matsumoto, Y. (2005), Information Extraction and Sentence Classification applied to Clinical Trial MEDLINE Abstracts, *in* 'Proceedings of the 2005 International Joint Conference of InCoB, AASBi and KSB', pp. 85–90.

Harkema, H., Roberts, I., Gaizauskas, R. & Hepple, M. (2005), Information Extraction from Clinical Records, *in* 'Proceedings of the 4th UK e-Science All Hands Meeting'.

Joachims, T. (1999), Making Large-scale SVM Learning Practical, *in* B. Schölkopf, C. Burges & A. Smola, eds, 'Advances in Kernel Methods-Support Vector Learning'.

Kudo, T. & Matsumoto, Y. (2004), A Boosting Algorithm for Classification of Semi-Structured Text, *in* 'EMNLP2004'.

Lin, J., Karakos, D., Demner-Fushman, D. & Khudanpur, S. (2006), Generative Content Models for Structural Analysis of Medical Abstracts, *in* 'Proceedings of the 2006 Workshop on Biomedical Natural Language Processing'.

McKnight, L. & Srinivasan, P. (2003), Categorization of Sentence Types in Medical Abstracts, *in* 'AMIA 2003 Symposium Proceedings', pp. 440–444.

Moschitti, A. (2004), A Study on Convolution Kernels for Shallow Semantic Parsing, *in* 'Proceedings of the 42-th Conference on Association for Computational Linguistic'.

Niu, Y. & Hirst, G. (2004), Analysis of Semantic Classes in Medical Text for Question Answering, *in* 'Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains'.

Niu, Y., Zhu, X., Li, J. & Hirst, G. (2005), Analysis of Polarity Information in Medical Text, *in* 'Proceedings of the American Medical Informatics Association 2005 Annual Symposium', pp. 570–574.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S. & Tsujii, J. (2005), Developing a Robust Part-of-Speech Tagger for Biomedical Text, *in* 'Advances in Informatics - 10th Panhellenic Conference on Informatics', pp. 382–392.

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, Springer.

Yamamoto, Y. & Takagi, T. (2005), A Sentence Classification System for Multi Biomedical Literature Summarization, *in* 'Proceedings of the 21st International Conference on Data Engineering'.