Annotating and Recognising Named Entities in Clinical Notes

Yefeng Wang
School of Information Technology
The University of Sydney
Australia 2006
ywang1@it.usyd.edu.au

Abstract

This paper presents ongoing research in clinical information extraction. This work introduces a new genre of text which are not well-written, noise prone, ungrammatical and with much cryptic content. A corpus of clinical progress notes drawn form an Intensive Care Service has been manually annotated with more than 15000 clinical named entities in 11 entity types. This paper reports on the challenges involved in creating the annotation schema, and recognising and annotating clinical named entities. The information extraction task has initially used two approaches: a rule based system and a machine learning system using Conditional Random Fields (CRF). Different features are investigated to assess the interaction of feature sets and the supervised learning approaches to establish the combination best suited to this data set. The rule based and CRF systems achieved an F-score of 64.12% and 81.48% respectively.

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 improve processes in the work of clinical wards, and to gain greater understanding of patient care as well as the progression of diseases. However in some clinical contexts these clinical notes, as written by a clinicians, are in a less structured and often minimal grammatical form with idiosyncratic and cryptic shorthand. Whilst there is increasing interest in the automatic extraction of the contents of clinical text, this particular type of notes cause significant difficulties for automatic extraction processes not present for well-written prose notes.

The first step to the extraction of structured information from these clinical notes is to achieve accurate identification of clinical concepts or named entities. An entity may refer to a concrete object mentioned in the notes. For example, there are 3 named entities - *CT*, *pituitary macroadenoma* and *suprasellar cisterns* in the sentence: CT revealed pituitary macroadenoma in suprasellar cisterns.

In recent years, the recognition of named entities from biomedical scientific literature has become the focus of much research, a large number of systems have been built to recognise, classify and map biomedical terms to ontologies. However, clinical terms such as findings, procedures and drugs have received less attention. Although different approaches have been proposed to identify clinical concepts and map them to terminologies (Aronson, 2001; Hazlehurst et al., 2005; Friedman et al., 2004; Jimeno et al., 2008), most of the approaches are language pattern based, which suffer from low recall. The low recall rate is mainly due to the incompleteness of medical lexicon and expressive use of alternative lexico-grammatical structures by the writers. However, only little work has used machine learning approaches, because no training data has been available, or the data are not available for clinical named entity identification.

There are semantically annotated corpora that have been developed in biomedical domain in the past few years, for example, the GENIA corpus of 2000 Medline abstracts has been annotated with biological entities (Kim et al., 2003); The PennBioIE corpus of 2300 Medline abstracts annotated with biomedical entities, part-of-speech tag and some Penn Treebank style syntactic structures (Mandel, 2006) and LLL05 challenge task corpus (Nédellec, 2005). However only a few corpora are available in the clinical domain. Many corpora are ad hoc annotations for evaluation, and

Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, pages 18–26,
Suntec, Singapore, 4 August 2009.

the size of the corpora are small which is not optimal for machine learning strategies. The lack of data is due to the difficulty of getting access to clinical text for research purposes and clinical information extraction is still a new area to explore. Many of the existing works focused only on clinical conditions or disease (Ogren et al., 2006; Pestian et al., 2007). The only corpus that is annotated with a variety of clinical named entities is the CLEF project (Roberts et al., 2007) .

Most of the works mentioned above are annotated on formal clinical reports and scientific literature abstracts, which generally conform to grammatical conventions of structure and readability. The CLEF data, annotated on clinical narrative reports, still uses formal clinical reports. The clinical notes presented in this work, is another genre of text, that is different from clinical reports, because they are not well-written. Notes written by clinicians and nurses are highly ungrammatical and noise prone, which creates issues in the quality of any text processing. Examples of problems arising from such texts are: firstly, variance in the representation of core medical concepts, whether unconsciously, such as typographical errors, or consciously, such as abbreviations and personal shorthand; secondly, the occurrences of different notations to signify the same concept. The clinical notes contain a great deal of formal terminology but used in an informal and unorderly manner, for example, a study of 5000 instances of Glasgow Coma Score (GCS) readings drawn from the corpus showed 321 patterns are used to denote the same concept and over 60% of them are only used once.

The clinical information extraction problem is addressed in this work by applying machine learning methods to a corpus annotated for clinical named entities. The data selection and annotation process is described in Section 3. The initial approaches to clinical concept identification using both a rule-based approach and machine learning approach are described in Section 4 and Section 5 respectively. A Conditional Random Fields based system was used to study and analyse the contribution of various feature types. The results and discussion are presented in Section 6.

2 Related Work

There is a great deal of research addressing concept identification and concept mapping issues. The Unified Medical Language System Metathesaurus (UMLS) (Lindberg et al., 1993) is the world's largest medical knowledge source and it has been the focus of much research. The simplest approaches to identifying medical concepts in text is to maintain a lexicon of all the entities of interest and to systematically search through that lexicon for all phrases of any length. This can be done efficiently by using an appropriate data structure such as a hash table. Systems that use string matching techniques include SAPHIRE (Hersh and Hickam, 1995), IndexFinder (Zou et al., 2003), NIP (Huang et al., 2005) and MaxMatcher (Zhou et al., 2006). With a large lexicon, high precision and acceptable recall were achieved by this approach in their experiments. However, using these approaches out of box for our task is not feasible, due to the high level of noise in the clinical notes, and the ad hoc variation of the terminology, will result in low precision and recall.

A more sophisticated and promising approach is to make use of shallow parsing to identify all noun phrases in a given text. The advantage of this approach is that the concepts that do not exist in the lexicon can be found. MedLEE (Friedman, 2000) is a system for information extraction in medical discharge summaries. This system uses a lexicon for recognising concept semantic classes, word qualifiers, phrases, and parses the text using its own grammar, and maps phrases to standard medical vocabularies for clinical findings and disease. The MetaMap (Aronson, 2001) program uses a three step process started by parsing free-text into simple noun phrases using the Specialist minimal commitment parser. Then the phrase variants are generated and mapping candidates are generated by looking at the UMLS source vocabulary. Then a scoring mechanism is used to evaluate the fit of each term from the source vocabulary, to reduce the potential matches (Brennan and Aronson, 2003). Unfortunately, the accurate identification of noun phrases is itself a difficult problem, especially for the clinical notes. The ICU clinical notes are highly ungrammatical and contain large number of sentence fragments and ad hoc terminology. Furthermore, highly stylised tokens of combinations of letters, digits and punctuation forming complex morphological tokens about clinical measurements in non-regular patterns add an extra load on morphological analysis, e.g. "4-6ml+/hr" means 4-6 millilitres or more secreted by

the patient per hour. Parsers trained on generic text and MEDLINE abstracts have vocabularies and language models that are inappropriate for such ungrammatical texts.

Among the state-of-art systems for concept identification and named entity recognition are those that utilize machine learning or statistical techniques. Machine learners are widely used in biomedical named entity recognition and have outperformed the rule based systems (Zhou et al., 2004; Tsai et al., 2006; Yoshida and Tsujii, 2007). These systems typically involve using many features, such as word morphology or surrounding context and also extensive post-processing. A state-of-the-art biomedical named entity recognizer uses lexical features, orthographic features, semantic features and syntactic features, such as part-of-speech and shallow parsing.

Many sequential labeling machine learners have been used for experimentation, for example, Hidden Markov Model(HMM) (Rabiner, 1989), Maximum Entropy Markov Model (MEMM) (McCallum et al., 2000) and Conditional Random Fields (CRF) (Lafferty et al., 2001). Conditional Random Fields have proven to be the best performing learner for this task. The benefit of using a machine learner is that it can utilise both the information form of the concepts themselves and the contextual information, and it is able to perform prediction without seeing the entire length of the concepts. The machine learning based systems are also good at concept disambiguation, in which a string of text may map to multiple concepts, and this is a difficult task for rule based approaches.

3 Annotation of Corpus

3.1 The Data

Data were selected form a 60 million token corpus of Royal Prince Alfred Hospital (RPAH)'s Intensive Care Service (ICS). The collection consists of clinical notes of over 12000 patients in a 6 year time span. It is composed of a variety of different types of notes, for example, patient admission notes, clinician notes, physiotherapy notes, echocardiogram reports, nursing notes, dietitian and operating theatre reports. The corpus for this study consists of 311 clinical notes drawn from patients who have stayed in ICS for more than 3 days, with most frequent causes of admission. The patients were identified in the patient records using keywords such as cardiac disease, liver disease, respiratory disease, cancer patient, patient underwent surgery etc. Notes vary in size, from 100 words to 500 words. Most of the notes consist of content such as chief complaint, patient background, current condition, history of present illness, laboratory test reports, medications, social history, impression and further plans. The variety of content in the notes ensures completely different classes of concepts are covered by the corpus. The notes were anonymised, patient-specific identifiers such as names, phone numbers, dates were replaced by a like value. All sensitive information was removed before annotation.

Category	Example
FINDING	*lung cancer*; *SOB*; *fever*
PROCEDURE	*chest X Ray*;*laparotomy*
SUBSTANCE	*Ceftriaxone*; CO_2; *platelet*
QUALIFIER	*left*; *right*;*elective*; *mild*
BODY	*renal artery*; *LAD*; *diaphragm*
BEHAVIOR	*smoker*; *heavy drinker*
ABNORMALITY	*tumor*; *lesion*; *granuloma*
ORGANISM	*HCV*; *proteus*; *B streptococcus*
OBJECT	*epidural pump; larnygoscope*
OCCUPATION	*cardiologist*; *psychiatrist*
OBSERVABLE	*GCS*; *blood pressure*

Table 1: Concept categories and examples.

3.2 Concept Category

Based on the advice of one doctor and one clinician/terminologist, eleven concept categories were defined in order to code the most frequently used clinical concepts in ICS. The eleven categories were derived from the SNOMED CT concept hierarchy. The categories and examples are listed in Table 1. Detailed explanation of these categories can be found in SNOMED CT Reference Guide[1]

3.3 Nested Concept

Nested concepts are concepts containing other concepts and are annotated in the corpus. They are of particular interest due to their compositional nature. For example, the term *left cavernous carotid aneurysm embolisation* is the outermost concept, which belongs to PROCEDURE. It contains several inner concepts: the QUALIFIER *left* and the term *cavernous carotid aneurysm* as a FINDING,

[1] SNOMED CT ® Technical Reference Guide - July 2008 International Release. http://www.ihtsdo.org/

which also contains *cavernous carotid* as BODY and *aneurysm* as ABNORMALITY.

The recognition of nested concepts is crucial for other tasks that depend on it, such as coreference resolution, relation extraction, and ontology construction, since nested structures implicitly contain relations that may help improve their correct recognition. The above outermost concept may be represented by embedded concepts and relationships as: *left cavernous carotid aneurysm embolisation* IS A *embolisation* which has LATERALITY *left*, has ASSOCIATED MORPHOLOGY *aneurysm* and has PROCEDURE SITE *cavernous carotid*.

3.4 Concept Frequency

The frequency of annotation for each concept category are detailed in Table 2. There are in total 15704 annotated concepts in the corpus, 12688 are outermost concepts and 3016 are inner concepts. The nested concepts account for 19.21% of all concepts in the corpus. The corpus has 46992 tokens, with 18907 tokens annotated as concepts, hence concept density is 40.23% of the tokens. This is higher than the density of the GENIA and MUC corpora. The 12688 annotated outermost concepts, results in an average length of 1.49 tokens per concept which is less than those of the GENIA and MUC corpora. These statistics suggest that ICU staff tend to use shorter terms but more extensively in their clinical notes which is in keeping with their principle of brevity.

The highest frequency concepts are FINDING, SUBSTANCE, PROCEDURE, QUALIFIER and BODY, which account 86.35% of data. The remaining 13.65% concepts are distributed into 6 rare categories. The inner concepts are mainly from QUALIFIER, BODY and ABNORMALITY, because most of the long and complex FINDING and PROCEDURE concepts contain BODY, ABNORMALITY and QUALIFIER, such as the example in Section 3.3.

3.5 Annotation Agreement

The corpus had been tokenised using a whitespace tokeniser. Each note was annotated by two annotators: the current author and a computational linguist experienced with medical texts. Annotation guidelines were developed jointly by the annotators and the clinicians. The guidelines were refined and the annotators were trained using an iterative process. At the end of each iteration, annotation agreement was calculated and the annotations were reviewed. The guidelines were modified if necessary. This process was stopped until the agreement reached a threshold. In total 30 clinical notes were used in the development of guidelines. Inter-Annotator Agreement (IAA) is reported as the F-score by holding one annotation as the standard. F-score is commonly used in information retrieval and information extraction evaluations, which calculates the harmonic mean of recall and precision as follows:

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

The IAA rate in the development cycle finally reached 89.83. The agreement rate between the two annotators for the whole corpus by exact matching was 88.12, including the 30 development notes. An exact match means both the boundaries and classes are exactly the same. The instances where the annotators did not agree were reviewed and relabeled by a third annotator to generate a single annotated gold standard corpus. The third annotator is used to ensure every concept is agreed on by at least two annotators.

Disagreements frequently occur at the boundaries of a term. Sometimes it is difficult to determine whether a modifier should be included in the concept: *massive medial defect* or *medial defect*, in which the latter one is a correct annotation and *massive* is a severity modifier. Mistakes in annotation also came from over annotation of a general term: *anterior approach*, which should not be annotated. Small disagreements were caused by ambiguities in the clinical notes: some medical

Category	Outer	Inner	All
ABNORMALITY	0	926	926
BODY	735	1331	2066
FINDING	4741	71	4812
HEALTHPROFILE	399	0	399
OBJECT	179	23	202
OBSERVABLE	198	227	425
OCCUPATION	139	0	139
ORGANISM	36	17	53
PROCEDURE	2353	39	2392
QUALIFIER	1659	21	1680
SUBSTANCE	2249	361	2610
TOTAL	12688	3016	15704

Table 2: Frequencies for nested and outermost concept.

devices (OBJECT) are often annotated as PROCEDURE, because the noun is used as a verb in the context. Another source of disagreement is due to the ambiguity in clinical knowledge: it was difficult to annotate the man-made tissues as BODY or SUBSTANCE, such as *bone graft* or *flap*.

4 Rule Based Concept Matcher

4.1 Proofreading the Corpus

Before any other processing, the first step was to resolve unknown tokens in the corpus. The unknown tokens are special orthographies or alphabetic words that do not exist in any dictionary, terminologies or gazetteers. Medical words were extracted from the UMLS lexicon and SNOMED CT (SNOMED International, 2009), and the MOBY (Ward, 1996) dictionary was used as the standard English word list. A list of abbreviations were compiled from various resources. The abbreviations in the terminology were extracted using pattern matching. Lists of abbreviations and shorthand were obtained from the hospital, and were manually compiled to resolve the meaning. Every alphabetic token was verified against the dictionary list, and classified into *Ordinary English Words*, *Medical Words*, *Abbreviations*, and *Unknown Words*.

An analysis of the corpus showed 31.8% of the total tokens are non-dictionary words, which contains 5% unknown alphabetic words. Most of these unknown alphabetic words are obvious spelling mistakes. The spelling errors were corrected using a spelling corrector trained on the 60 million token corpus, Abbreviations and shorthand were expanded, for example *defib* expands to *defibrillator*. Table 3 shows some unknown tokens and their resolutions. The proofreading require considerable amount of human effort to build the dictionaries.

unknown word	examples	resolution
CORRECT WORD	bibasally	bibasally
MISSING SPACE	oliclinomel	Oli Clinomel
SPELLING ERROR	dolaseteron	dolasetron
ACRONYM	BP	blood pressure
ABBREVIATION	N+V	Nausea and vomiting
SHORTHAND	h'serous	haemoserous
MEASUREMENT	e4v1m6	GCS measurement
SLASHWORDS	abg/ck/tropt	ABG CK Tropt
READINGS	7mg/hr	

Table 3: Unknown tokens and their resolutions.

4.2 Lexicon look-up Token Matcher

The lexicon look-up performed exact matching between the concepts in the SNOMED CT terminology and the concepts in the notes. A hash table data structure was implemented to index lexical items in the terminology. This is an extension to the algorithm described in (Patrick et al., 2006). A token matching matrix run through the sentence to find all candidate matches in the sentence to the lexicon, including exact longest matches, partial matches, and overlapping between matches. Then a Viterbi algorithm was used to find the best sequence of non-overlapping concepts in a sentence that maximise the total similarity score. This method matches the term as it appears in the terminology so is not robust against term variations that have not been seen in the terminology, which results in an extremely low recall. In addition, the precision may be affected by ambiguous terms or nested terms.

The exact lexicon look-up is likely to fail on matching long and complex terms, as clinicians do not necessarily write the modifier of a concept in a strict order, and some descriptors are omitted. for example *white blood cell count normal* can be written as *normal white cell count*. In order to increase recall, partial matching is implemented. The partial matching tries to match the best sequence, but penalise non-matching gaps between two terms. The above example will be found using partial matching.

5 CRF based Clinical Named Entity Recogniser

5.1 Conditional Random Fields

The concept identification task has been formulated as a named entity recognition task, which can be thought of as a sequential labeling problem: each word is a token in a sequence to be assigned a label, for example, B-FINDING, I-FINDING, B-PROCEDURE, I-PROCEDURE, B-SUBSTANCE, I-SUBSTANCE and so on. Conditional Random Fields (CRF) are undirected statistical graphical models, which is a linear chain of Maximum Entropy Models that evaluate the conditional probability on a sequence of states give a sequence of observations. Such models are suitable for sequence analysis. CRFs has been applied to the task

of recognition of biomedical named entities and have outperformed other machine learning models. CRF++[2] is used for conditional random fields learning.

5.2 Features for the Learner

This section describes the various features used in the CRF model. Annotated concepts were converted into BIO notation, and feature vectors were generated for each token.

Orthographic Features: Word formation was genaralised into orthographic classes. The present model uses 7 orthographic features to indicate whether the words are captialised or upper case, whether they are alphanumeric or contains any slashes, as many findings consist of captialised words; substances are followed by dosage, which can be captured by the orthography. Word prefixes and suffixes of character length 4 were also used as features, because some procedures, substances and findings have special affixes, which are very distinguishable from ordinary words.

Lexical Features: Every token in the training data was used as a feature. Alphabetic words in the training data were converted to lowercase, spelling errors detected in proofreading stage were replaced by the correct resolution. Shorthand and abbreviations were expanded into bag of words (*bow*) features. The left and right lexical bigrams were also used as a feature, however it only yielded a slight improvement in performance. To utilise the context information, neighboring words in the window $[-2, +2]$ are also added as features. Context window size of 2 is chosen because it yields the best performance. The target and previous labels are also used as features, and had been shown to be very effective.

Semantic Features: The output from the lexical-lookup system was used as features in the CRF model. The identified concepts were added to the feature set as semantic features, because the terminology can provide semantic knowledge to the learner such as the category information of the term. Moreover, many partially matched concepts from lexicon-lookup were counted as incorrectly matching, however they are single term head nouns which are effective features in NER.

Syntactic features were not used in this experiment as the texts have only a little grammatical structure. Most of the texts appeared in fragmentary sentences or single word or phrase bullet point format, which is difficult for generic parsers to work with correctly.

[2] http://crfpp.sourceforge.net/

Experiment	P	R	F-score
no pruning	58.76	26.63	36.35
exact matching	69.48	37.70	48.88
+proofreading	74.81	52.42	61.65
+partial matching	69.39	59.60	64.12

Table 4: Lexical lookup Performance.

6 Evaluation

This section presents experiment results for both the rule-based system and machine learning based system. Only the 12688 outermost concepts are used in the experiments, because nested terms result in multi-label for a single token. Since there is no outermost concepts in ABNORMALITY, the classification was done on the remaining 10 categories. The performances were evaluated in terms of recall, precision and F-score.

6.1 Token Matcher Performance

The lexical lookup performance is evaluated on the whole corpus. The first system uses only exact matching without any pre-processing of the lexicon. The second experiment uses a pruned terminology with ambiguous categories and unnecessary categories removed, but without proofreading of the corpus. The concept will be removed if it belongs to a category that is not used in the annotation. The third experiment used the proofreaded corpus with all abbreviations annotated. The fourth experiment was conducted on the proofread corpus allowing both exact matching and partial matching. The results are outlined in Table 4.

The lexicon lookup without pruning the terminologies achieved low precision and extremely low recall. This is mainly due to the ambiguous terms in the lexicon. By removing unrelated terms and categories in the lexicon, both precision and recall improved dramatically. Proofreading, correcting a large number of unknown tokens such as spelling errors or irregular conventions further increased both precision and recall. The 14.72 gain in recall mainly came from resolution and expansion of shorthand, abbreviations, and acronyms in the notes. This also suggest that this kind of clinical notes are very noisy, and require a consider-

able amount of effort in pre-processing. Allowing partial matching increased recall by 7.18, but decreased precision by 5.52, and gave the overall increase of 2.47 F-score. Partial matching discovered a larger number of matching candidates using a looser matching criteria, therefore decreased in precision with compensation of an increase in recall.

The highest precision achieved by exact matching is 74.81, confirming that the lexical lookup method is an effective means of identifying clinical concepts. However, it requires extensive effort on pre-processing both corpus and the terminology and is not easily adapted to other corpora. The lexical matching fails to identify long terms and has difficult in term disambiguation. The low recall is caused by incompleteness of the terminology. However, the benefit of using lexicon lookup is that the system is able to assign a concept identifier to the identified concept if available.

6.2 CRF Feature Performance

The CRF system has been evaluated using 10-fold cross validation on the data set. The evaluation was performed using the CoNLL shared task evaluation script [3].

The CRF classifier experiment results are shown in Table 5. A baseline system was built using only *bag-of-word* features from the training corpus. A context-window size of 2 and tag prediction of previous token were used in all experiments. Without using any contextual features the performance was 48.04% F-score. The baseline performance of 71.16% F-score outperformed the lexical-look up performance. Clearly the contextual information surrounding the concepts gives a strong contribution in identification of concepts, while lexical-lookup hardly uses any contextual information.

The full system is built using all features described in Section 5.2, and achieved the best result of 81.48% F-score. This is a significant improvement of 10.32% F-score over the baseline system. Further experimental analysis of the contribution of feature types was conducted by removing each feature type from the full system. $-$*bow* means bag-of-word features are removed from the full system. The results show only *bow* and *lexical-lookup* features make significant contribution to the system, which are 5.49% and 4.40% separately. *Bigram*, *orthographic*, *affixes* and *abbreviation* features each makes around $\sim$ 1% contribution to the F-score, which is individually insignificant, however the combination of them makes a significant contribution, which is 4.83% F-score.

Experiment	P	R	F-score
baseline	76.86	66.26	71.16
+*lexical-lookup*	82.61	74.88	78.55
full	**84.22**	**78.90**	**81.48**
$-$*bow*	81.26	73.32	77.08
$-$*bigram*	83.17	78.74	80.89
$-$*abbreviation*	83.20	77.26	80.12
$-$*orthographic*	83.67	78.24	80.87
$-$*affixes*	83.16	77.01	79.97
$-$*lexical-lookup*	79.06	73.15	75.99

Table 5: Experiment on Feature Contribution for the ICU corpus.

The most effective feature in the system is the output from the lexical lookup system. Another experiment using only *bow* and *lexical-lookup* features showed a boost of 7.39% F-score. This is proof of the hypothesis that using terminology information in the machine learner would increase recall. In this corpus, about one third of the concepts has a frequency of only 1, from which the learner as unable to learn anything from the training data. The gain in performance is due to the ingestion of semantic domain knowledge which is provided by the terminology. This knowledge is useful for determining the correct boundary of a concept as well as the classification of the concept.

6.3 Detailed CRF Performance

The detailed results of the CRF system are shown in Table 6. Precision, Recall and F-score for each class are reported. There is a consistent gap between Recall and Precision across all categories. The best performing classes are among the most frequent categories. This is an indication that sufficient training data is a crucial factor in achieving high performance. SUBSTANCE, PROCEDURE and FINDING are the best three categories due to their high frequency in the corpus. However, QUALIFIER achieved a lower F-score because qualifiers usually appear at the boundaries of two concepts, which is a source of error in boundary recognition.

Low frequency categories generally achieved high precision and low recall. The recall decreases as the number of training instances decreases, be-

[3] http://www.cnts.ua.ac.be/conll2002/ner/bin/

Class	P	R	F-score
BODY	72.00	64.29	67.92
FINDING	83.17	78.74	80.89
BEHAVIOR	83.87	72.22	77.61
OBJECT	75.00	27.27	40.00
OBSERVABLE	89.47	56.67	69.39
ORGANISM	0.00	0.00	0.00
PROCEDURE	87.63	81.09	84.24
QUALIFIER	75.80	75.32	75.56
OCCUPATION	87.50	41.18	56.00
SUBSTANCE	91.90	88.53	90.19

Table 6: Detailed Performance of the CRF system.

cause there is not enough information in the training data to learn the class profiles. It is a challenge to boost the recall of rare categories due to the variability of the terms in the notes. It is not likely that the term would match to the terminology, and hence there would be no utilisation of the semantic information.

Another factor that causes recognition errors is the nested concepts. BODY achieved the least precision because of the high frequency of nested concepts in its category. The nested construction also causes boundary detection problems, for example *C5/6 cervical discectomy* PROCEDURE is annotated as *C5/6* BODY and *cervical discectomy* PROCEDURE.

The results presented here are higher than those reported in biomedical NER system. Although it is difficult to compare with other work because of the different data set, but this task might be easier due to the shorter length of the concepts and fewer long concepts (avg. 1.49 in this corpus vs. avg. 1.70 token per concept in GENIA). Local features would be able to capture most of the useful information while not introducing ambiguity.

7 Future Work and Conclusion

This paper presents a study of identification of concepts in progressive clinical notes, which is another genre of text that hasn't been studied to date. This is the first step towards information extraction of free text clinical notes and knowledge representation of patient cases. Now that the corpus has been annotated with coarse grained concept categories in a reference terminology, a possible improvement of the annotation is to reevaluate the concept categories and create fine grained categories by dividing top categories into smaller classes along the terminology's hierarchy. For example, the FINDING class can be further divided into SYMPTOM/SIGN, DISORDER and EVALUATION RESULTS. The aim would be to achieve better consistency, less ambiguity and greater coverage of the concepts in the corpus.

The nested concepts model the relations between atomic concepts within the outermost concepts. These structures represent important relationships within this type of clinical concept. The next piece of work could be the study of these relationships. They can be extended to represent relationships between clinical concepts and allow for representing new concepts using structured information. The annotation of relations is under development. The future work will move from concept identification to relation identification and automatic ontology extension.

Preliminary experiments in clinical named entity recognition using both rule-based and machine learning approaches were performed on this corpus. These experiments have achieved promising results and show that rule based lexicon lookup, with considerable effort on pre-processing and lexical verification, can significantly improve performance over a simple exact matching process. However, a machine learning system can achieve good results by simply adapting features from biomedical NER systems, and produced a meaningful baseline for future research. A direction to improve the recogniser is to add more syntactic features and semantic features by using dependency parsers and exploiting the unlabeled 60 million token corpus.

In conclusion, this paper described a new annotated corpus in the clinical domain and presented initial approaches to clinical named entity recognition. It has demonstrated that practical acceptable named entity recognizer can be trained on the corpus with an F-score of 81.48%. The challenge in this task is to increase recall and identify rare entity classes as well as resolve ambiguities introduced by nested concepts. The results should be improved by using extensive knowledge resource or by increasing the size and improving the quality of the corpus.

Acknowledgments

The author wish to thank the staff of the Royal Prince Alfred Hospital, Sydney : Dr. Stephen Crawshaw, Dr. Robert Herks and Dr Angela Ryan

for their support in this project.

References

R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *In Proceeding of the AMIA Symposium*,17–21.

F. Brennan and A. Aronson 2003. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. *Journal of Biomedical Informatics*,36(4/5),334–341.

A. Côté and American Veterinary Medical Association and College of American Pathologists. 2009. Snomed International. *College of American Pathologists.*

C. Friedman 2000. A broad coverage natural language processing system. *In Proceedings of the AMIA Symposium*,270–274.

C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. 2004. Automated Encoding of Clinical Documents Based on Natural Language Processing. *Journal of the American Medical Informatics Association*,11(5),392–402.

B. Hazlehurst, R. Frost, F. Sittig, and J. Stevens. 2005. MediClass: A System for Detecting and Classifying Encounter-based Clinical Events in Any Electronic Medical Record. *Journal of the American Medical Informatics Association*,12(5),517–529.

R. Hersh, and D. Hickam. 1995. Information retrieval in medicine: The SAPHIRE experience. *Journal of the American Society for Information Science*,46(10),743–747.

Y. Huang, J. Lowe, D. Klein, and J. Cucina. 2005. Improved Identification of Noun Phrases in Clinical Radiology Reports Using a High-Performance Statistical Natural Language Parser Augmented with the UMLS Specialist Lexicon. *Journal of the American Medical Informatics Association*,12(3),275–285.

A. Jimeno, et al. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*,9(3).

D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Journal of Bioinformatics*, 19(1),180–182.

J. Lafferty et al. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data *Machine learning-international workshop then conference*, 282–289.

A. Lindberg et al. 1993. The Unified Medical Language System. *Methods Inf Med.*

M. Mandel 2006. Integrated Annotation of Biomedical Text: Creating the PennBioIE corpus. *Text Mining Ontologies and Natural Language Processing in Biomedicine*, Manchester, UK.

A. McCallum, et al. 2000. Maximum entropy Markov models for information extraction and segmentation *Proc. 17th International Conf. on Machine Learning*, 591–598.

C. N'edellec. 2005. Learning Language in Logic - Genic Interaction Extraction Challenge. *Proceedings of the ICML05 Workshop on Learning Language in Logic*, Bonn, 31–37.

V. Ogren, G. Savova, D. Buntrock, and G. Chute. 2006. Building and Evaluating Annotated Corpora for Medical NLP Systems. *AMIA Annu Symp Proceeding.*.

J. Patrick, Y. Wang, and P. Budd. 2006. Automatic Mapping Clinical Notes to Medical Terminologies *In Proceedings of Australasian Language Technology Workshop*.

P. Pestian, C. Brew, P. Matykiewicz, J. Hovermale, N. Johnson, K. Cohen, and W. Duch. 2007. A Shared Task Involving Multi-label Classification of Clinical Free Text *In Proceedings of BioNLP workshop*.

R. Rabiner 1989. A tutorial on hidden Markov models and selected applications inspeech recognition *Proceedings of the IEEE*,77(2), 257–286.

A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, A. Setzer, I. and Roberts. 2007. The CLEF Corpus: Semantic Annotation of Clinical Text. *AMIA Annu Symp Proceeding.*, Oct 11:625–629.

R. Tsai, L. Sung, J. Dai, C. Hung, Y. Sung, and L. Hsu. 2006. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity *BMC Bioinformatics*.

G. Ward 1996. Moby thesaurus. http://etext.icewire.com/moby/.

K. Yoshida, and J. Tsujii. 2007. Reranking for Biomedical Named-Entity Recognition *Proceedings of BioNLP 2007: Biological, translational, and clinical language processing*, 209–216.

G. Zhou, J. Zhang, J. Su, D. Shen, and L. Tan. 2004. Recognizing Names in Biomedical Texts: a Machine Learning Approach *BioInformatics*, 20(7) 1178–1190.

X. Zhou, et al. 2006. MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup. *Proc PRICAI*,1145–1149.

Q. Zou. 2003. IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. *Proc AMIA Symp*,763–767.