Cascading Classifiers for Named Entity Recognition in Clinical Notes

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

Jon Patrick
School of Information Technology
University of Sydney
Australia
jonpat@it.usyd.edu.au

Abstract

Clinical named entities convey great deal of knowledge in clinical notes. This paper investigates named entity recognition from clinical notes using machine learning approaches. We present a cascading system that uses a Conditional Random Fields model, a Support Vector Machine and a Maximum Entropy to reclassify the identified entities in order to reduce misclassification. Voting strategy was employed to determine the class of the recognised entities between the three classifiers. The experiments were conducted on a corpus of 311 manually annotated admission summaries form an Intensive Care Unit. The recognition of 10 types of clinical named entities using 10 fold cross-validation achieved an overall results of 83.3 F-score. The reclassifier effectively increased the performance over stand-alone CRF models by 3.35 F-score.

Keywords

Named Entity Recognition, Clinical Information Extraction, Machine Learning, Classifier Ensemble, Two Phase Model

1 Introduction

With the rapid growth of clinical data produced by health organisations, efficient information extraction from these free text clinical notes will be valuable for improving the work of clinical wards and gaining greater understanding of patient care as well as progression of disease. Recognising named entities is a key to unlocking the information stored in unstructured clinical text. Named entity recognition is an important subtask of Information Extraction. It involves the recognition of named entity (NE) phrases, and usually the classification of these NEs into particular categories. In the clinical domain, important entity categories are clinical findings, procedures and drugs.

In recent years, the recognition of named entities in the biomedical scientific literature has become the focus of much research. A large number of systems have been built to recognise, classify and map biomedical entities to ontologies. On the other side, only a little work have been reported in clinical named entity recognition [14, 8, 17]. NER has achieved high performance in scientific articles and newswire text, whereas the clinical notes written by clinicians are in a less structured and often minimal grammatical form with idosyncratic and cryptic shorthand, which poses challenges in NER. Principally, the clinical named entity recognition systems are rule or pattern based. The rules or patterns may not be generalisable due to the specific writing style of individual clinicians. However, a machine learning approach is not fully advanced in clinical named entity recognition due to a lack of available training data. We have investigated the issues of clinical named entity recognition, by constructing a set of annotation guidelines and manually annotating 311 clinical notes from an Intensive Care Unit (ICU), with inter-annotator agreement of 88%. In this paper we present a named entity recogniser using a cascade of classifiers to find entities. The named entities will serve as a prerequisite for clinical relation extraction, clinical notes indexing and question answering from the ICU database.

There have been many approaches to NER in biomedical literature. They roughly fall into three approaches: rule-based approaches, dictionary-based approaches and machine learning based approaches. The state-of-art machine learning based systems focus on selecting effective features for building classifiers. Many machine learners have been used for experimentation, for example, Support Vector Machines (SVMs)[9], Hidden Markov Model (HMM)[16], Maximum Entropy Model (ME) [2] and Conditional Random Fields (CRFs) [12]. Conditional Random Fields have been proven to be the best performing learner for the NER task [3]. The benefit of using a machine learner is that it can utilise both the information form of the entity themselves and the contextual information surrounding the entity. It has better generalisability over pattern based approach as it is able to perform prediction without seeing the entire length of the entity.

Nevertheless the performance of biomedical NER systems still trails behind newswire NER systems. It suggests that individual NER system may not cover entity representations with sufficiently rich features due to the great variety and ambiguity in biomedical named entities. This problem also exists in clinical text as it has characteristic of both formal and informal linguistic styles, with many unseen named entities, spelling variations and abbreviations. To overcome these difficulties, we propose a classifier cascade approach to clinical NER. We firstly build a CRF based

Workshop Biomedical Information Extraction 2009 - Borovets, Bulgaria, pages 42–49

classifier to identify the boundary and class of the named entities, then we trained a SVM and an ME model to reclassify the class of the named entities using the output of the CRF models and different features. The final class of the entity was determined by a majority voting [18] among the output of the CRF, SVM and ME models. The overall system achieved best performance of 83.26 F-score. The cascading classifiers improved 3.35 F-sore over the stand-alone CRF system.

This paper is organised as follows: Section 2 gives an overview of related work in biomedical named entity recognition. Section 3 introduces the data used in our experiments. Section 4 to Section 6 describes the cascading named entity recogniser in detail. Section 7 presents the evaluation of the proposed system as well as discussion of the results.

2 Related Work

The early research in biomedical named entity recogniton was dictionary based. The Unified Medical Language System Metathesaurus (UMLS) is the word's largest medical knowledge source and it has been widely used as the dictionary for identification of medical named entities in clinical reports. Systems such as [23, 22, 7] use string matching methods to find UMLS concepts in clinical notes. These systems suffer low recall due to the great variety in medical terminology. A more sophisticated 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 named entities that do not exist in the dictionary can be found. For example, MedLEE [6] and MetaMap [1] program utilised parsers to parse text into noun phrases then map these phases to standard medical vocabularies. However, accurate identification of noun phrases is itself a problem. Most parsers trained on formal medical text or newswire articles may not be directly applicable to ungrammatical clinical text.

Among the state-of-art systems for biomedical named entity recognition are those that utilise machine learning approach [19, 5, 21]. Machine learning approaches have been successfully applied in biomedical named entity recognition and outperformed rule-based systems. With an annotated corpus, the machine learner is able to learn models to make prediction on unseen data. Recent research has found that using stand alone machine learners may not be enough for biomedical named entity recognition due to the complex structure of the named entity. Most of the learners only use local information about the current word, while correct identification of many named entities requires global information over the entire entity. To employ global information into the learner, rule based post-processing or using multiple classifiers is required.

Cascading of classifiers has become a new research direction in machine learning recently. It can effectively improve performance of individual classifiers. The combination of the results of different classifiers is able to overcome possible local weakness of individual classifiers and produce more reliable recognition results. Many of the current named entity recognition systems use a classifier combination strategy such as [13, 11, 20, 3, 4]. For example, Lee et al. [13] divide NER into recognition and classification, and employed two SVMs for recognition and classification. Kim et al [11], uses a similar two phase approach to separate recognition from classification. In their system, CRF was used to identify the named entity boundaries and SVMs are used for assigning entity categories. Chan et al. [3] further extended the two phase model using CRFs for both boundary identification and entity classification. On the other hand, cascading systems also achieved promising results. Yoshida et al. [20] uses an ME classifier to produce the n-best tag sequences for the input text and uses a ME-based log-linear classifier to find the best sequence. The combination of models effectively increased the performance by 1.55 F-score on the GENIA corpus [10]. Similarly, Corbett and Copestake [4] use an ME classifier and an ME rescorer in recognising chemical named entities from chemistry papers, the cascading approach gives about a 3 point increase in F-score over the stand alone system.

Entity Class	**Example**	*n*
FINDING	*lung cancer*; *SOB*;	4741
PROCEDURE	*chest X Ray*;*laparotomy*	2353
SUBSTANCE	*Ceftriaxone*; CO_2; *platelet*	2449
QUALIFIER	*left*; *right*; *elective*; *mild*	2353
BODY	*renal artery*; *liver*	735
BEHAVIOR	*smoker*; *heavy drinker*	399
ORGANISM	*HCV*; *proteus*	36
OBJECT	*pump; larnygoscope*	179
OCCUPATION	*cardiologist*; *psychiatrist*	139
OBSERVABLE	*GCS*; *blood pressure*	192

Table 1: *Named Entity classes with examples and number of instances in the corpus.*

3 The Data

We have developed a set of annotation guidelines for clinical named entities and manually annotated 311 admission summaries from an hospital's Intensive Care Unit (ICU). The clinical notes were drawn from patients who have stayed in ICU for more than 3 days, with the most frequent causes of admission such as cardiac disease, liver disease, respiratory disease, cancer patient, patient underwent surgery and so on. 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. Notes are de-identified before annotation.

The guidelines were developed using an iterative approach. The clinicians and linguists jointly defined the annotation schema. The entity classes are mainly based on the SNOMED CT concept categories, and SNOMED CT user development guide[1]. The guidelines defined 10 entity types, which are detailed in Table 1. Firstly, the clinicians and linguists jointly annotated 10 notes and produced initial guidelines. The guidelines were then refined using five iterations

[1] http://www.ihtsdo.org/publications/

Class	P	R	F
OVERALL	89.22	87.05	**88.12**
BODY	87.40	82.48	84.87
OBSERVABLE	84.77	79.52	82.06
QUALIFIER	89.89	81.80	85.66
OBJECT	78.35	80.00	79.17
SUBSTANCE	95.01	94.03	94.52
BEHAVIOUR	80.49	78.57	79.52
OCCUPATIONS	78.95	77.92	78.43
FINDING	91.72	91.17	91.44
ORGANISM	75.00	70.59	72.73
PROCEDURE	87.43	87.82	87.63

Table 2: *The inter-annotator agreement measured by F-score for 10 Entity Classes.*

of annotation and analysis. Five notes were used in each iteration, at the end of each cycle, the clinicians and linguists discussed the disagreements and made amendment to the guidelines if necessary. Finally the development annotation agreement reached a stable state and the guidelines were finalised.

The remainder of the annotation was completed by 2 computational linguists with medical knowledge and experience in biomedical NLP. During annotation, the annotators constantly consulted the domain experts from the hospital. Most of the clinical text can be understood by the linguists even though they do not have a clinical background. The meaning of most terms can be determined by the linguistic constructs of the text. Some difficult terms require a dictionary lookup to resolve the meaning. A few abbreviations are not easily understood by the clinicians either, so they needed to check the abbreviation lists to identify the terms. The polysemous abbreviations sometimes cause mistakes in annotations, but for most of the cases their meaning can be resolved by looking at the context.

The inter-annotator agreement was found to be 88% F-score and the agreement of each individual category is presented in Table 2, which indicates the upper bound of the NER performance. The two annotators have similar backgrounds, therefore their annotation is relatively consistent when applying the guidelines. Most of the entities were annotated using their linguistic knowledge rather than clinical knowledge. However the annotation guidelines also specified some clinical information that required domain knowledge. For example, the causation of a clinical symptom or a particular drug used to treat a certain disease. It was not easy for computational linguists to discover this knowledge as there are no explicit rules to define them. Thus the true recall of the annotation will be lower than the annotation created by clinicians.

4 Methods

We built a named entity recognition system using a cascade of classifiers. The first component in the system is a CRF based model. It is similar to most of the stand-alone named entity recognition systems, that integrated a set of features to produce a sequence of named entity labels. Then a reclassifier is built using different feature sets with the output of the CRF model aimed at reclassifying misclassified named entities produced by CRF model. The system architecture is illustrated in Figure 1. We experimented with two different machine learning models ME and SVMs in the reclassification stage. The output of these two models are then combined with the output of the CRF model to produce a final class for the named entity.

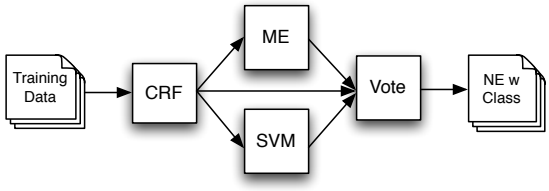

Fig. 1: *System architecture of CRF model with reclassifiers.*

The named entity recognition task has been formulated as a sequence labeling task. The named entities are represented in BIO notation, where B denotes the beginning of an entity, I denotes inside, but not at beginning of an entity and O denotes not in any part of a entity. Each word is a token in an input sequence to be assigned a label. The output is a sequence of BIO tags. for example, B-FINDING, I-FINDING, B-PROCEDURE, I-PROCEDURE and so on. Figure 2 presents a sentence annotated with BIO tags.

Head $_{\text{B-PROCEDURE}}$ CT $_{\text{I-PROCEDURE}}$ revealed $_{\text{O}}$ pituitary $_{\text{B-FINDING}}$ macroadenoma $_{\text{I-FINDING}}$ in $_{\text{O}}$ suprasellar $_{\text{B-BODY}}$ cisterns $_{\text{I-BODY}}$. $_{\text{O}}$

Fig. 2: *An example sentence with BIO tags.*

5 CRF-based Named Entity Recogniser

Conditional Random Field (CRF) is a discriminative probabilistic model that is useful for the labeling sequential data. It aims to maximize the conditional probability of the output given an input sequence. The CRFs have several advantages over ME, SVM and HMM in sequential labeling tasks. It can use the sequential information where the output is the most likely tag sequence over the entire input sequence, whereas SVMs and ME don't consider sequence information. Modeling conditional probability rather than joint probability does not suffer from strong Markov assumptions on the input and output sequence distributions of HMMs. Because of these two properties, CRFs have an advantage over other learners and have been shown to be useful in biomedical named entity recognition in previous work [3].

5.1 Features for CRF Learner

Word Features: Every token in the training data was used as a feature. Alphabetic words in the training data were converted to lowercase in order to in-

crease recall. The left and right lexical bigrams were also used as a feature, however it only yielded a slight improvement in performance.

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, for example many findings consist of captialised words and whether they are alphanumeric or contain any slashes.

Affixes: prefixes and suffixes of character length 4 were also used as features, because some procedures, substances and findings have special affixes, which are quite distinguishable from ordinary words.

Context Features: To utilise the context information, neighboring words within a context window size of 5 are added as features, i.e. two previous tokens and two next tokens. Window size of 5 is chosen because it yields the best performance. The target and previous entity class labels are also used as features, and had been shown to be very effective.

Dictionary Features: We constructed two different features to capture the existence of an entity in a closed dictionary and an open dictionary. The closed dictionary is constructed by extracting all entity names from the training data in each fold. The open dictionary was constructed from SNOMED CT terminology. Single word concepts and the rightmost head nouns of multi-word concepts were extracted. The category was assigned to the word when it is used as a feature. For words belonging to more than one class, all the classes were represented in the feature. For example the word *aspiration* was found in both the finding and procedure dictionaries, the feature is represent as Open/Procedure/Finding. The open dictionary consists of 25468 entries.

Abbreviations and Acronyms: The abbreviation lists were constructed from 3 resources: abbreviations from SNOMED CT terminology, abbreviations & acronyms from the hospital and manually resolved abbreviations in the larger corpus. We constructed the SNOMED CT lists using rules to extract abbreviations and acronyms from the gloss of SNOMED concepts, for example, *AAA - Abdominal aortic aneurysm (disorder)* is extracted as a pair of abbreviations and expanded. We also obtained a list of commonly used abbreviations from the intensive care unit's database. The corpus abbreviation list was obtained by first using orphographical and lexical patterns to extract a list of candidate abbreviations from a larger collection of notes that the training data were drawn from. The extracted candidates were then manually verified by two human experts.

When a word is matched to an abbreviation, the class of the abbreviation is assigned to the word as a feature. Moreover, the two rightmost words in the expansion are used as a feature. The abbreviation lists consists of 9757 entries. However, building abbreviation lists requires a great deal of manual work.

POS Features: The POS tags of 3 words surrounding the target words (1 preceding and 2 following) are considered. The POS features is able to generalise the low frequency words. The use of POS helps to determine the boundaries of named entities. The experiments conducted by Zhou and Su [21] discovered POS features are very useful in biomedical NER. The POS tagger used to generate POS tags is the GENIA tagger[2]. This is a tagger trained on biomedical abstracts. It is not expected the tagger will produce high accuracy tagging results on our corpus, but the POS is relatively simple syntactic processing, and might be useful.

6 Reclassifier

The re-classifier aims to reclassify the semantic categories of the named entities recognised by the CRF learner. As we observed there are many misclassifications produced by the CRFs because the local context of different named entity classes are similar.

6.1 The Learning Algorithms

We experimented with MEs and SVMs for reclassification. SVM is a supervised machine learner based on the theory of structural risk minimization, which aims to find an optimal hyperplan to separate the training example into two classes, and make predictions based on these support vectors. SVMs have been successfully applied to many NLP tasks such as document classification. It can use large numbers of features and dose not make the feature independence assumption. The SVMs are binary classifiers so we use one-vs-the-rest approach for multi-label classification and choose the final prediction based on the smallest margin to the hyperplane.

The Maximum Entropy (ME) model is a probabilistic machine learner that models the conditional probability of output o for given inputs history h. The conditional probability is defined as:

$$P(o|h) = \frac{1}{Z_\lambda(h)} exp\left(\sum_{i=1}^{k} \lambda_i f_i(h,o)\right)$$

where $f_i(h,o)$ is a binary-valued feature function, λ_i is the weighting parameter of $f_i(h,o), k$ is the number of features and $Z_\lambda(h)$ is a normalisation factor for $\sum_0 p(o|h) = 1$.

6.2 Features for Reclassifier

Word Unigram: Words described in CRF features were mainly adapted in reclassifier. The words inside the entity were used as bag of words features, i.e. we didn't consider the order and position of the word. However, the position of words are important. The class of the entities are usually determined by the head noun of the phrase, for example the head noun *pain* in *chest pain* and *abdominal pain* determines the class of these entities. These head nouns are usually at the right most position of a named entity. We also consider words at the rightmost position of the entity and the second rightmost word as entity context features.

Word Bigram: The word bigrams inside the entities were used as features. For example, the bigram of the entity "chronic renal failure" is "chronic renal" and "renal failure".

[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

Word Trigram: The word trigrams inside the entities were used as a feature.

Orthography: Orthographic features described in Section 5.1 were used.

Context Words: The 2 words to the left boundary of the entities and the 2 words to the right boundary of the entities were used as context features.

Character n-grams: The character n-grams of each word in the entity were used as features. character 3-grams and character 4-grams were used as features. It is observed that some of the clinical named entities are derived from latin, that have special prefix, suffixes or substrings. For example procedures often end with *-tomy*, some diseases end with *-itis*, and some drug names have special substrings.

Dictionary Features: We use the same dictionary list, but we made 2 different feature types: The non-positional words, which is the same as Dictionary Features used in CRF model; and Positional, where only the last word in the entities were matched to the dictionary.

Abbreviation Features: The abbreviation list is the same as that used in CRF features. The class of the abbreviation for the matched word is used as a feature, however we also expand the matched abbreviation and use the words in expansion as the bag of word features. For example, the entity CRF is expanded to Chronic Renal Failure and all three words in the expansion are used as features. All the words in an abbreviation with more than one expansion were used as a bag-of-words, such as LAD is expanded into "left axis deviation" and "left anterior descending artery". All seven words are used as bag-of-word features. A binary feature is used to indicate if the expansion is unique, the value set to 0 if there is only one expansion for the abbreviation.

CRF Output Class: The class predicted by the CRF model was used as a feature in reclassifiers.

6.3 Training the Reclassifier

We divided the training set into 5 folds and use 4 folds to train a CRF model and make prediction on the remaining fold. The remaining fold is used to generate training data for reclassifiers. We repeat the process 5 times, each time holding out a different fold as test set, until all instances in the training set have the the CRF predicted class value. The reclassifiers were trained using all data generated by this procedure. This procedure makes sure the reclassifier is not trained on the output of the CRFs that is trained on the data need to be classified by the reclassifier.

6.4 Voting for Reclassification

We use a voting method for the re-classifier ensemble. This ensemble strategy uses heuristic rules to judge which results to be selected if the individual learners cannot reach a consensus decision. We use a majority vote strategy to decide the final class. The class prediction produced by the CRF model was used in voting between the output of CRF, ME and SVMs. The final class is assigned if two of the learners agree. If the three classifiers produce three different outputs, the results were ranked by the probability produced by the CRF, ME and SVM models. The probability of SVMs were obtained by converting the distance between the instance and hyper-plane produced by the SVM using an sigmoid function [15]. The probability of CRFs were obtained by the highest probability of the tag in the entity tag sequence. Although the probabilities are all between 0 and 1, however, one flaw in the probability ranking is that different classifiers use different weight functions, so some probabilities may not be directly comparable. An adjusted probability function should be learnt from the corpus.

6.5 Separating Recognition from Classification

We separate the entity recognition from entity classification. The system structure is illustrated in Figure 3. The CRF model was used to identify the boundaries of the named entities. The entity labels were converted to B-ENT and I-ENT if the phrase is an entity. After the recognition stage, the identified entities were sent to the ME and SVMs reclassifiers for identification of the class of the entity. The outputs of ME and SVMs were used for voting using the method described in Section 6.4.

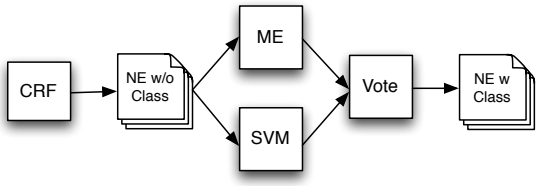

Fig. 3: *System architecture for separating recognition from classification.*

7 Experimental Results

7.1 Experimental Setup

The data consists of a total of 45953 tokens, 17544 tokens are annotated with entity tags. The tag density is 38.18%. There are in total 12882 named entities results with an average of 1.36 tokens per named entity. The results were evaluated by 10-fold cross-validation. Each fold was stratified on a sentence level, so that for the rare classes such as ORGANISM had some instances in each fold. We adapted the evaluation scripts provided by the JNLPBA 2004 shared task to evaluate the system performance[3]. The standard Recall/Precision/F-score are used as evaluation metrics.

We use CRF++[4] package for CRF learning. CRF++ takes the standard CoNLL NER shared task input. We converted the data and features into the accepted format and trained the model using the package's default parameter configuration. We did no feature selection and all folds use the same parameter setting. CRF++ can produce output tags along with

[3] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html
[4] http://crfpp.sourceforge.net/

the tag's probability, these probabilities are used for reclassification.

We use LibSVM[5] and Maxent[6] for reclassification. We use the polynomial kernel with degree 2 in SVM learning, and set the C values to 8, i.e. approximately the ratio of the number of negative instances to the number of positive instances in the training data. The other parameters are obtained by a 10-fold cross-validation on the training data. The probability of SVM tags are obtained by setting appropriate software options to enable probability output during training and prediction. To train the Maxent model, we use Maxent package's default parameters and terminate the learning process when the training model converges.

7.2 CRF Classifier Performance

Table 3 shows the performance of the CRF classifier. Features were added to the model progressively to understand the contribution of each feature. The overall performance is very promising, with a score F-score of 79.91. All experiments used window size of 5 and previously predicted labels. The baseline model was built using only word features. The dictionary features are very useful, the use of a dictionary allows for the identification of unseen words in the test set. The dictionary entries also act as trigger words described in some biomedical NER systems, and can help identify the boundary of entity. POS tag is not as effective as expected, this may be due to the inaccurate POS tagging by the GENIA tagger and that the sentences are poorly structured. Other features all make moderate contribution to the performance. Different context window sizes were investigated and a window size 5 produced the best performance.

Feature Sets	P	R	F
Word	79.82	66.28	72.41
+Orthographic	77.96	71.37	74.52
+Affix	78.24	72.59	75.31
+Dictionary	82.77	75.76	79.11
+Abbreviation	83.19	76.38	79.64
+POS	83.30	76.78	**79.91**
window size 0	69.82	56.28	63.32
window size 3	82.74	75.23	78.80
window size 5	83.30	**76.78**	**79.91**
window size 7	**83.63**	74.57	78.84

Table 3: *Contribution of features by adding features progressively (using window size of 5). Different window sizes were investigated.*

7.3 Reclassifier Performance

We built the reclassifiers using the output of the 5-fold cross trained CRF output. Table 4 shows the performance of the reclassifiers on the test data. We compared SVM reclassifier performance with ME reclassifier performance. The SVM and ME have the same level of performance on classification, with SVM slightly outperforming ME by about 0.4 F-score. The classification performance is high, which suggests that if the boundary of a named entity is correctly identified, the performance of the NER will go above 90 F-score, and identifying boundaries is more difficult than assigning named entity classes.

Class	SVM P/R/F	ME P/R/F
overall	93.20/93.20/93.20	92.81/92.81/92.81
body	86.85/75.35/80.60	85.78/76.73/80.92
finding	91.30/95.62/93.41	90.62/95.73/93.10
hprofile	94.39/88.22/90.96	95.87/86.84/90.98
object	92.50/55.36/68.43	88.00/47.82/60.91
obs.	94.32/80.17/86.16	91.87/80.79/85.36
organism	55.56/22.22/31.48	50.00/19.00/27.38
procedure	93.82/91.24/92.49	93.91/90.42/92.12
qualifier	99.62/97.83/98.72	99.68/97.91/98.79
social	94.33/81.90/86.50	96.33/74.04/83.03
substance	93.58/96.48/95.00	93.15/95.60/94.36

Table 4: *Results of reclassification for correctly identified named entities.*

7.4 Cascading System Performance

The reclassifiers were run on the CRF output to correct misclassified labels. The overall performance of the cascade system were evaluated. We also evaluated the performance of separating recognition from classification. In recognition, the CRF models only predict whether or not a phrase is an entity.

Table 5 shows the performance of the cascading classifiers. *CRF only* is the baseline model without reclassification. CRF recognition reports the entity boundary performance by CRF. The rest are reclassification results with SVM, ME and Voting respectively. In general the cascading systems outperform the stand alone CRF system. The performances vary from 2.03 to 3.35. This suggests that selecting different features for classification can further utilise the discriminative power of individual classifiers. The best combined system was obtained by using cascading classifiers with voting, which gives in total 3.35 F-score increase over the baseline CRF model. Cascading classifiers perform slightly better than recognition with reclassification, because recognition with reclassification cannot use the class information produced by the CRF model.

We trained two CRF models: the first one only uses 3 entity labels, B-ENT, I-ENT and O, and the sec-

System	P	R	F
CRF only	83.30	76.78	79.91
cascading SVM	85.42	80.69	82.99
cascading ME	85.02	80.31	82.60
cascading Voting	**85.87**	80.81	**83.26**
recognition + SVM	82.75	**82.16**	82.45
recognition + ME	82.28	81.69	81.98
recognition + Voting	84.65	80.99	82.78
CRF recognition 1	86.70	**86.08**	**86.39**
CRF recognition 2	**88.89**	83.90	86.32

Table 5: *Performance of combined systems using reclassification.*

[5] http://www.csie.ntu.edu.tw/ cjlin/libsvm/
[6] http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Class	CRF P/R/F	Cascading P/R/F
overall	83.30/76.78/79.91	85.87/80.81/83.26
body	74.67/59.57/66.21	75.89/66.39/70.83
finding	79.74/80.14/79.94	84.22/82.49/83.35
hprofile	86.85/67.15/75.73	86.34/74.44/79.95
object	81.67/23.82/35.60	71.70/42.46/53.33
obs.	82.04/57.77/67.78	79.61/63.02/70.35
organism	00.00/0.00/0.00	85.00/47.22/60.71
procedure	83.69/71.93/77.36	85.81/79.13/82.33
qualifier	88.19/85.34/86.72	87.62/86.62/87.12
social	74.83/26.55/39.01	73.61/38.13/50.24
substance	88.94/85.71/87.25	92.11/87.71/89.86

Table 6: *The performance of the best cascading system and baseline CRF systems with detailed information for each class.*

ond one uses all 21 entity tags. The first model produced a recognition performance of 86.70/86.08/86.39 in P/R/F. The recognition performance of the second model was obtained by changing all entity tags to B-ENT and I-ENT on the prediction output, which is 88.89/83.90/86.32 in P/R/F. The first model has higher recall than the second model, which results in higher recall in the Recognition with Reclassification model.

We use a different feature set in the Reclassifier from the CRF model because some features are not very informative in the CRF model, for example, adding the abbreviation expansion gives about 0.3 drop in F-score, and incorporating character-n gram features results in huge amount of features, which slows down the CRF learning process but results in insignificant [7] performance change.

7.5 Individual Class Performance

Table 6 shows the performance of overall cascading classifiers. We compared the best performing cascading system with the baseline CRF system. Overall, there is a consistent gap between precision and recall, with recall value 5 points F-score behind precision. The best-performing classes are among the most frequent classes. SUBSTANCE, FINDING and PROCEDURE are the best three categories due to their high frequency in the corpus. This is an indication that sufficient training data is a crucial factor in achieving both high precision and recall. BODY achieved the least accuracy among frequent classes. It is mainly caused by nested construction of the entities. Body entities can appear inside a nested entity at different positions for example, *chest* in *chest pain* and *ventricle* in *dilated ventricle*.

The low recall is caused by a lack of lexical information for named entities. In the corpus, about one third of the entities has a frequency of only one. To recognise these low frequency entities, generalised features are required to predict unseen examples. POS features and context features can partially cure this problem, but the lexical information is still being missed during the classification. The medical terminology has a great variety in its spelling plus clinicians invent new terms by combining morphologies during writing, such as inventing the term *rehaperisation*. It is difficult to capture unseen examples in test data for this small size corpus. Utilisation of external resources such as dictionary and abbreviation lists has shown its effectiveness in tackling this problem, but the external resources are not exhaustive and may not cover the dialect language used in different hospitals and clinical specialisations.

Reclassifiers use a great deal of word level features such as character n-grams that are focused on predicting labels of named entities, which effectively increased the performance by 3.5 point F-score. Reclassification increases the recall of infrequent classes. The CRF is likely to bias to the majority classes. Most of these rare class instances were classified as FINDING. Using more discriminative features Reclassifiers are able to separate these rare classes from majority classes. It has been shown that the SVM outperformed ME reclassifier. Combining the classification results of ME, SVMs and CRFs via voting has some positive influence on results, but not significant. The features used in the models are the same, which may cause correlation in misclassifications produced by the classifiers. The results might be improved using different feature sets for each learner, but the space for improvement is small. There is still around 3 points F-score in misclassification which maybe caused by human annotation errors.

Named Entity	CRF	RC	GS
frontal cavernoma	Body	Finding	Finding
E/O lesion	Proc.	Finding	Proc.
ST elevation	Proc.	Finding	Finding
smoker	Finding	H.profile	H.profile
CT Surg Reg	Proc.	Occup.	Occup.
Mac. larnygoscope	Finding	Proc.	Object
subclavian CVC	Finding	Object	Object
hilum	Body	Finding	Body
Tonsilectomy	Substa.	Proc.	Proc.

Table 7: *Some examples of classification disagreements between CRFs and Reclassifiers.*

We present some classification disagreements between the three classifiers in Table 7. RC indicates the reclassification results and GS is the gold-standard class. It is observed that the misclassifications appear more frequently in entities involved in abbreviations, ostensibly due to a lack of knowledge to resolve them. The reclassifiers make false correction at the rate of about 15%. The CRF is more likely to classify unseen entities into major categories whereas reclassifiers tend to classify the names according the the head nouns. The reclassifiers are biased to SVM and ME classifiers as the two learners used similar features for learning. There are about 20% entities assigned to different classes by each of the three classifiers.

The boundary detection achieved an F-score of 86.39. This performance is lower than the classification performance of $92 \sim 93$ F-score. Table 8 lists the partial matching performance of the system. As suggested by the results, many mistakes occurred at the boundary of the entities. Many of them are caused by the ambiguous modifiers at the boundaries of the phrase. Misrecognition in coordination structure is

[7] t-test 95% confidence interval

Matching Criteria	P	R	F
Exact Matching	85.87	80.81	83.26
Left Boundary	88.07	82.88	85.40
Right Boundary	89.77	84.48	87.05
Partial Matching	91.97	86.55	89.18

Table 8: *Results of different partial matching criteria.*

also a source of boundary error. This was demonstrated by the lower performance of BODY class, as they usually appear at the boundary of coordinated phases such as in *LAD and LCX stenosis*. Further investigation of recognition errors revealed several annotation errors. Inconsistent annotation of modifiers is a common mistake, for examples *medial defect* was annotated as *massive medial defect*, where the former is the correct annotation.

The overall result of the named entity recognition is promising, with only 5 points F-score behind the annotation agreement. Even with such noisy clinical text the system still reached an F-score of 83.26. The clinical named entities are relatively shorter in comparison to the biological named entity. Clinicians tend to use short terms and dense terminology in keeping with their principle of brevity. With the average length of only 1.36 tokens per entity, CRFs using contextual information are able to capture a significant portion of entity boundaries. The reclassifier uses global information about the entire term effectively to make corrections to misclassified entities.

8 Conclusion

We have presented a machine learning approach to clinical named entity recognition using a combination of machine learners. The system incorporated various features, and experimented with different strategies for combining machine learners. The cascading approach with voting among different classifier outputs produced the best results. With an improvement of 3.35 F-score from the baseline stand alone CRF classifier, the system achieved an overall result of 83.26 F-score. The performance gain is due to utilisation of global information of the entire entity to make correct predictions about misclassified entities. The future work will be focused on improving the boundary identification performance and injecting more domain knowledge into the named entity recognition system.

References

[1] A. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[2] A. Berger, V. Della Pietra, and S. Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[3] S. Chan and W. Lam. Efficient Methods for Biomedical Named Entity Recognition. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pages 729–735, 2007.

[4] P. Corbett and A. Copestake. Cascaded classifiers for confidence-based chemical named entity recognition. *BMC bioinformatics*, 9(Suppl 11):S4, 2008.

[5] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair. Exploiting context for biomedical entity recognition: From syntax to the web. In *Joint Workshop on Natural Language Processing in Biomedicine and Its Applications at Coling 2004*, 2004.

[6] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 2004.

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

[8] A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(Suppl 3):S3, 2008.

[9] T. Joachims, C. Nedellec, and C. Rouveirol. Text categorization with support vector machines: learning with many relevant. In *Machine Learning: ECML-98 10th European Conference on Machine Learning, Chemnitz, Germany*. Springer, 1998.

[10] J. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus-a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(90001):180–182, 2003.

[11] S. Kim, J. Yoon, K. Park, and H. Rim. Two-phase biomedical named entity recognition using a hybrid method. *Proceedings of The Second International Joint Conference on Natural Language Processing (IJCNLP-05)*, 3651:646–657, 2005.

[12] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 282–289, 2001.

[13] K. Lee, Y. Hwang, and H. Rim. Two-phase biomedical NE recognition based on SVMs. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*, pages 33–40. Association for Computational Linguistics Morristown, NJ, USA, 2003.

[14] P. Ogren, G. Savova, and C. Chute. Constructing evaluation corpora for automated clinical named entity recognition. In *Proc LREC*, 2008.

[15] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 1999.

[16] L. Rabiner et al. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[17] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, A. Setzer, and I. Roberts. Semantic annotation of clinical text: The CLEF corpus. In *Proceedings of Building and evaluating resources for biomedical text mining: workshop at LREC*, 2008.

[18] D. Ruta and B. Gabrys. Classifier selection for majority voting. *Information fusion*, 6(1):63–81, 2005.

[19] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA)*, pages 104–107, 2004.

[20] K. Yoshida and J. Tsujii. Reranking for Biomedical Named-Entity Recognition. *BioNLP 2007*, pages 209–216, 2006.

[21] G. Zhou and J. Su. Exploring deep knowledge resources in biomedical name recognition. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its applications (JNLPBA-2004)*, volume 171, 2004.

[22] X. Zhou, X. Zhang, and X. Hu. MaxMatcher: Biological concept extraction using approximate dictionary lookup. *In Proceeding of PRICAI*, pages 1145–1149, 2006.

[23] Q. Zou, W. Chu, C. Morioka, G. Leazer, and H. Kangarloo. IndexFinder: a method of extracting key concepts from clinical texts for indexing. In *AMIA... Annual Symposium proceedings [electronic resource]*, volume 2003, page 763. American Medical Informatics Association, 2003.