Biomedical Named Entity Recognition System

Jon Patrick and Yefeng Wang
Sydney Language Technology Research Group
School of Information Technologies
University of Sydney
{jonpat, ywang1@it.usyd.edu.au}

Abstract *We propose a machine learning approach, using a Maximum Entropy (ME) model to construct a Named Entity Recognition (NER) classifier to retrieve biomedical names from texts. In experiments, we utilize a blend of various linguistic features incorporated into the ME model to assign class labels and location within an entity sequence, and a post-processing strategy for corrections to sequences of tags to produce a state of the art solution. The experimental results on the GENIA corpus achieved an F-score of 68.2% for semantic classification of 23 categories and achieved F-score of 78.1% on identification.*

Keywords Named Entity Recognition, ME model, Information Retrieval.

1 Introduction

The discovery of the human gene and rapid developments in the biomedical domain has produced large amounts of genetic data. This has resulted in exponential growth of biomedical literature over the past few years. MEDLINE, the primary research database serving the biomedical community, currently contains over 14 million abstracts, with 60,000 new abstracts appearing each month. This growth of biomedical literature has given rise to a pressing need for automatic information extraction from the data bank.

Biomedical literature contains a rich set of biomedical entities providing key information to access the knowledge. A biomedical named entity is a word or sequence of words that can be classified as a name or biomedical term, such as protein, DNA, RNA, etc. Named Entity Recognition is the task of identifying and semantically classifying named entities in text. In the biomedical domain, the goal of the biomedical named entity recognition (BioNER) task is to find the biomedical terms such as names of genes, proteins, gene products, organisms, drugs, chemical compounds etc. in texts and classify them into their correct categories. It is a critical step for future automatic processing of biomedical literature to be mounted on a large scale, and further to perform high level biomedical information extraction task such as analysis and question answering.

BioNER consists of two tasks, term identification and term classification. Identification finds the region of a named entity in a text. Its main goal is to differentiate between terms and non-terms without looking at the semantic meaning of a term. However term classification determines the semantic concept of that named entity and assigns it to a biomedical class, such as genes, proteins or DNA.

The named entity recognition in the newswire domain has been studied for a long time and has achieved 90% accuracy [11]. However, named entity recognition in biomedical domain has different characteristics, with an accuracy of only around 70%. Despite the "near human" performance of named entity recognition in newswire domain, many similar strategies do not work well when adapted into the biomedical domain because of the distinctive nature of this task Hirschman et al., [5] Tuason et al.,[15] Shen et al.,[9] Lin et al., [8] and Lee et al.,[6].

First, biomedical named entities are not conventional proper nouns. They are usually unknown words containing uncommon orthographic features such as hyphens, digits, letters, and Greek letters. Furthermore, there are no conventional rules for biomedical term formation.

Second, biomedical terms may have a number of spelling variations. For example, the term *Alpha UF1 cells* may have spelling variations: *Alpha UF-1 cells*, *Uf-1 Alpha cell.* Such variations always cause recognition ambiguity.

Third, ambiguity and inconsistency are often encountered in named entity classification. Many named entities with the same orthographical features may fall into different categories, for example, nested entities of one category may contain an NE of another category, or a NE is composed of two NEs from different categories.

Fourth, complex naming and abbreviation

Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.

conventions can differ from organism to organism, and class to class. Abbreviations tend to be a short form and coincide with English words such as “can”, “dot”. In addition, the abbreviations are intrinsically degenerate forms, so that one abbreviation can have a number of meanings, depending on the document domain.

Fifth, new named entities are introduced daily as new substances are discovered and some existing terms might change as our understanding changes. The system must be able to recognize new names and unseen names, and this causes difficulties in rule based systems.

In this paper, we explore machine learning (ML) and natural language processing (NLP) techniques to recognize biomedical named entities in text. We present a strategy that is different to previous work on two bases, firstly we use a framework that incorporates as many useful linguistic features as possible for this recognition task, and secondly we use a Maximum Entropy (ME) model as the basis of our machine learning system, finally we apply rule-based post-processing on the classification results.

2 Related work

Named Entity Recognition in the biomedical domain is more difficult than in a newswire domain because of the complex name formation of the biomedical terms and our current lack of experience in understanding optimal strategies to solve this task. Current NER approaches include: dictionary based, rule based, machine learning based, and hybrid approach. Due to the spelling variation and complex naming convention of biomedical terms, NER systems that rely on dictionary resources and pre-built rules do not seem to perform well, especially for large scale tasks.

2.1 Dictionary and rule based approaches

Early approaches in biomedical named entity recognition typically were dictionary-based approaches and rule based approaches. These approaches use domain specific heuristic rules and rely heavily on existing dictionaries, representative research includes Krauthammer et al.,[19]; Hirschman et al., [5]; Tuason [15]. However, the dictionary-based approaches typically perform quite poorly, with coverage generally only in the range of 10-30%, even allowing for some variability in the form of names. The rule-based systems perform well for existing named entities, but they usually perform poorly on new named entities and it is costly to adapt them to new entity classes. Once a new class is introduced, a set of new rules has to be generated manually. Since there is no standard biomedical term naming convention, the rule building process becomes more difficult as the number of class increases. Furthermore, the rule-based system performs poorly on larger copra, Gaizauskas et al., [4] and Fukuda et al., [3].

2.2 Machine-learning approaches

The major problem in machine learning based NER systems is the lack of training data. Before the GENIA corpus 3.0, Kim et al., [20] there was no consistent annotated corpus, so researchers used some small-scale data sets, such as GENIA 1.1 and Bio1. The development of GENIA 3.0 which contains 2000 abstracts provides a standard evaluation data set for the machine learning approach. Many other corpora that derived from the GENIA corpus have been constructed, such as the BioNLP/NLPBA corpus.

The typical machine learning algorithms include Naive Bayes (NB), Support Vector Machine (SVM), Hidden Markov Model (HMM), Maximum Entropy (ME) models, and Conditional Random Fields (CRF). Kazama et al. [6] used an SVM to achieve an F-score of 54.4 on GENIA 1.1. Nobata and Collier [1] incorporated rich features into a hidden Markov Model and achieved an F-score of 75.9 on a primary version of GENIA, which contains 100 medical abstracts, Shen and colleague [9] further enhanced the HMM model by exploring some special phenomena and a rule based postprocessor. They have achieved performance of 66.5 on the GENIA 3.0 corpus. Lin and colleagues [8] adapted a maximum entropy model for biomedical named entity recognition with a post processor, and achieved the performance of an F-score of 72.1. Finally CRF have been introduced into this field. Settles [10], Tsai et al.,[13] and shown good results. (69.9% and 69.8%) on JNLPBA corpus.

A large body of post processing has been proposed for biomedical named-entity recognition, typical work includes Shen et al., [9], Lin et al., [8], Zhou et al., [17]. Shen et al. proposed a rule based system for cascaded named entity resolution. They automatically extract rules from the training corpus. Lin et al., make use of a rule based boundary extension strategy combined with dictionary lookup for reclassification, and this post-processing effectively increases performance by about 20%.

3 Modelling the data

3.1 GENIA corpus

The GENIA corpus is an annotated corpus of paper abstracts extracted from the MEDLINE database using the MeSH query, *human*, *blood cell* and *transcription factor*. In the current version 3.02, 2000 abstracts are annotated by domain experts with entity tags. The annotation of the biomedical terms is based on the GENIA ontology. The GENIA ontology is a taxonomy of 48 biologically relevant categories. In our system, we recognize 23 distinct entity classes, including Protein, OtherName, DNA, CellType, CellLine, OtherOrganicCompound, Lipid, MultiCell, Virus, RNA, Tissue, CellComponent, Peptide, BodyPart, AminoAcidMonomer, OtherArtificial-Source, Polynucleotide, MonoCell, Atom, Inorganic, Nucleotide, and Carbohydrate.

3.2 Maximum entropy machine learner

In our experiments, we adapt Zhang Le's Maximum Entropy Tool Kit[1]. This is a C++ implementation of OpenMaxent, which contains Generalized Iterative Scaling (GIS) parameter estimation and Gaussian Prior Smoothing algorithms.

One advantage of using a maximum entropy model is that the features need not be statistically independent, and therefore it is easy to incorporate features with dependencies. Some of the features used in this system are strongly dependent, and yet they do not bias the ME model overly much, thus the ME models can yield better probability estimates compared with some other probability based machine learners, such as Hidden Markov Model (HMM) and Naïve Bayes classifier. Another advantage of using ME model is that it is scalable and does not suffer from the data sparseness problem. The training speed of ME model is faster than SVM. Although the training time is a one-time cost in a real word application, however, in prototyping a system, training must be fast enough to allow experimentation with various configurations.

We use a simple bag of word model to represent the language model of the texts hence each token's features are represented by a binary attribute value.

We employ the simplest BIO representation which is widely used in named entity recognition tasks, for example, Kazama et al. [6]. B means the token is at beginning of an NE, I means the token is in an NE, and O means the token is not in a named entity. For each category C, we have B_C and I_C tags to represent the beginning and inside of an NE of that category.

4 Feature set

Machine Learning systems typically represent data in terms of a set of features. It is intended that these feature sets encode the most significant aspects of the data for the learning task. In this section, we describe the features we used in our system.

4.1 Orthographical features

Orthographical features are used to capture the rendition of words, such as capitalization, digitalisation and punctuation. Orthographic features allow strings to be compared based on their spelling characteristics and are widely used in the biomedical and newswire domains, such as Shen et al., Collier et al., and Tsai et al.,[9,2,13].

Table 1 presents some orthographic feature used in our system. The feature such as AllCaps, for words with only capital letters, is useful to identify biomedical abbreviations. The CapsAndDigits feature is a very strong indicator of entities from Protein, DNA and Othername classes. The comma, colon, bracket, full stop and stop words are useful for detecting the boundaries of named entities. The Greek letters and Roman numerals are often used in biomedical terms, and the feature of LowercaseOnly strongly indicates the non-entity class.

Features	Example
AllCap	ALAS, HIV, RIP
SingleCap	B, M, T
DigitNumbers	7, 8 , 41
CapsAndDigit	CD4, MEK1
InitCapDigit	Am80
TwoCaps	FcR, FasL
InitCapsLowcase	Ras, Crkl, Ctx,
InitCaps	FURa
LowCapsMix	dNTPs, dPRL,
LowcaseOnly	protein, cell
LetterAndDigit	ETh1, h1RaK
InitDigit	15B7, 17q, 1A9
Backslash	/
Parenthesis	[,], (,)
Punctuations	;,:,,,.
Hypen	-
RomanNumeral	I, II, III
HasHyphen	-induced, Eth-1
GreekLetters	Alpha, kappa
Other	Other symbols

Table 1. Orthographic Features with examples.

4.2 Part of speech feature

In the newswire domain, the POS features have been shown to be of limited use because the POS features may adversely interact with the use of some important capitalization information [14]. However, POS features are widely used in the biomedical domain [9,3,13,16], because many biomedical entities are in lowercase, and capitalization information in the biomedical domain is not as evidential as that in the newswire domain. Moreover, since the biomedical named entities have many elements, identifying the boundaries is a more difficult task. The POS tagging can help to determine the boundaries, as for example, verbs and prepositions usually indicate a boundary.

[1] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

In our experiments, each word is assigned a POS tag feature. The GENIA POS tagger Tsuruoka et al. [14] was used to provide the POS information. The GENIA POS tagger is specifically tuned for biomedical text such as the MEDLINE abstracts, which reported 98.20% accuracy on the GENIA corpus.

4.3 Affix features

The prefix and suffix can provide good clues for classifying named entities, and has been widely used in Kazama et al.,[6] Zhou et al.,[17] Tsai et al.,[13] and Lee et al.,[6]. Kazama et al. collected the 10,000 most frequent prefixes/suffixes from the training data while Zhou et al. construct a prefix/suffix list using a statistical method and grouped the prefix/suffix into 23 categories using a weighted score according to the prefix/suffix distributions.

We extracted the affixes from each class of the corpus by their diversity and frequency. Frequent and diverse affixes may have higher priority to be extracted, for example, the suffix *~cyte* is usually a cell type and the suffix *~lipid* is usually a lipid. However, short affixes always conflict with common English words, for example, the suffix *~ase* conflicts with English word "disease". Some common affixes have high diversity and frequency in both entity and non-entity classes, so they do not contribute to the classification, for example, the suffix *~tion*.

We extracted the 3500 most frequent prefixes and suffixes from the training data, we filter the prefixes and suffixes if the root-diversity is less than 5 (examples in Table 2).

Suffix	**Class**	**Example**
~nase	Protein	Kinase
~hift	Othername	Shift
~esis	Othername	embryogenesis
~ytes	CellType	leukocytes
~ycin	OtherOrganicComp.	rapamycin
~eria	MonoCell	Bacteria
STAT~	Protein	STAT1s
NFAT~	Protein	NFAT2
path~	Othername	Pathogenic

Table 2. Examples of Suffix Features extracted from the corpus.

4.4 Unigram named entity feature

The unigram term is similar to the core-term proposed by Fukuda et al., [3] and the single term list in Lee's system [6]. It is a list consisting of all single word named entities, such as *IL-2*, *NF-kappaB*. These terms usually have special surface clues, and appear at the leftmost part of an NE. They can be combined with a head noun to form a new named entity. We extract all unigram named entities from the training corpus, and remove them if their frequency is less than 5.

4.5 Head noun feature

The head noun is usually the major element of a noun phrase, which describes the function or the property of the named entity. For example, the *NF-kappaB activation* is the head noun for the named entity *CoCl2-induced NF-kappaB activation*. Some previous works Nobata et al., [18] and Shen et al., [9] show that the head nouns in biomedical named entities can provide significant clues for distinguishing the entity classes. For example, the term *IL-6 kappa B binding factor* is classified as a Protein, and the *L-6 kappa B motif* is classified as DNA. Hence, the classification is determined by the head nouns *binding factor* and *motif.*

We constructed a head noun list by first looking at the rightmost word in a named entity, since the head nouns usually are the last noun in the named entities. A list of head noun candidates was extracted from each named entity class and ranked by frequency, because the most frequent nouns can be a good predictor for that class. We filter out the head nouns with frequency less than 5. Table 3 lists some head nouns extracted from the training data.

Class	**Head nouns**
Protein	factor, protein, receptor, complex, heterodimer, subunit, kinases, calcineurin, selectin, antibody
Other Name	expression, activity, activation, differentiation, apoptosis, phosphorylation, production, assays, levels
DNA	promoter, site, gene, element, chromosome, plasmid, repeat, construct, locus
Cell Type	Lymphocyte, monocyte, macrophage, neutrophils

Table 3. Examples of Head Noun Features

4.6 Bi-gram phrase feature

We extracted all bi-gram noun phrases from the entities from the training corpus as a feature. We filter the low frequency bi-gram phrases, as we found they cause some negative effects. The bi-gram phrase is similar to the bi-gram head nouns, except we included some high frequency word bi-grams and bi-gram

T cell	transcription factor
gene expression	cell line
virus type	human monocytes
signal transduction	Epstein-Barr virus

Table 4. Examples of bi-gram phrase features.

named entities. Table 4 lists some high frequency bi-gram phrases.

4.7 Contextual information

The contextual information is important for this task. The words preceding and following the target words are also used as features in our experiments.

5 Post-processing

5.1 Fixing inconsistent tag sequence

As we used the B, I, O notation to indicate the location of the token within the NE, the system may produce an inconsistent class sequence such as "*O B_Protein I_DNA O*". However, only a consistent sequence of tags is annotated as a named entity. We identified four types of such inconsistency in classifications, and describe rules using regular expressions to fix these mistakes.

1. I tag without preceding B tags. These tags are mainly due to false positives and partially identified terms. Some lower case words that have been seen in the named entities are classified as *I_OtherName*. We change this type of invalid I tags into O tags as we assumed that fixing these I tags can increase recall to a certain degree. Further inconsistencies of a sequence of I tags is altered so that the first is a B tag.

2. Missing middle I tag. Some middle I tags are classified as O tags, such as "and", "or". We change these O tags according to the preceding B class or I class tag.

3. Inconsistent I tag sequence. The I tag sequence in some long entities may be mixed with I tags from another class, for example, "*O B_DNA I_Protein I_DNA O*". We fix this mistake by changing the inconsistent I tag class into the B tag class.

4. Inconsistent tag sequence due to nested named entity. We found in our experiments, many entities are tagged as "*O B_C1 I_C1 I_C2 O*", where *I_C1* and *I_C2* are tags from two different categories. The I_C2 is usually a head noun and "*B_C1 I_C1*" is a NE from C1. We fix this inconsistency by first checking if I_C2 is in the head noun list, and then assign the NE a class according the head noun's category.

5.2 Rule based boundary correction

We found a number of partially identified named entities are due to missing the rightmost head nouns or the leftmost adjectives. We built from the training data a list of head nouns and a list of modifiers that frequently appear in the boundaries of a NE. Then we designed two simple rules to perform the boundary correction which is similar to the boundary extension in Lin et al. [8].

1. NE := NE + headnoun
2. NE := modifier + NE

After the entity recognition is completed by our ME-model and keeping the tag sequences fixed, we applied these two rules on recognized named entities to expand the boundary to the right and left.

6 Experiments and discussion

To conduct experiments, we divided the 2000 abstracts into a training set and test set. The training set consisted of 1800 abstracts and the test set consisted of 200 abstracts. The performance was measured by precision, recall and F-score, which are the standard measures for named entity recognition. The accuracy is measured by the number of correctly recognized named entities.

The main computational cost of the ME model is the GIS parameter estimation, which involves computation of each observed expectation, and re-computation of the model's expectation on each iteration. The greater the number of iterations the better the training accuracy. Since the number of iterations we need for the model to converge to an optimal solution is unknown, we ran 2000 iterations for each experiment. The experiment settings are shown in Table 5

Training (#words)	Testing (#words)	Context (position)	Iteration
415,761	43,597	-2,-1,0,1,2	2,000

Table 5. Experiment configuration.

6.1 The contribution of features

The task was to investigate the contribution of linguistic features to predicting the correct class boundaries and labels. Several experiments were performed using different combinations of features. (Results in Table 6)

The orthographic features (O) are only

	Feature	P	R	F	Effect
1	O	0.331	0.197	0.247	
2	O+P	0.408	0.317	0.357	0.110
3	O+P+HN	0.584	0.549	0.566	0.209
4	O+P+HN+UE	0.611	0.571	0.590	0.024
5	O+P+UE+HN+A	0.625	0.589	0.606	0.016
6	O+P+UE+HN+BP	**0.626**	**0.596**	**0.611**	0.020
7	O+P+UE+HN+BP+A	0.616	0.585	0.600	-0.011
8	O+P+UE+HN+ALLBP	0.625	0.588	0.606	-0.005

Table 6. The contribution of features and the progressive effects from adding more features.

moderately informative, only 4 NE categories are recognized in any way. NEs among the minor categories cannot be identified and most of the entities are classified as Protein, as most have the same surface appearance as protein names. The overall F-score achieved is 0.247. Addition of the POS features (P) provides limited information on the classifications. It leads to an increase of 0.110 on the F-score. The Head noun feature (HN) is very useful and provides a positive effect of 0.209 on the F-score compared to using simple Orthography plus POS tags. Adding in the unigram entity feature (UE) also provides a further improvement of 0.024 in F-value. These four features (O+P+HN+UE) are the most informative features, so if we treat these four features (Exp 4) as a new baseline for the remainder of the experiments we can discuss each other experiment relative to this baseline.

The Affix features (A) lead to a small positive improvement by 0.016 (Exp 5). Adding the bi-gram phrase feature (BP) (Exp 6) gives a slight increase in F-value (0.020). However combining affix features and bi-gram phrase features together (Exp 7) slightly degrades the performance by 0.011 and 0.006 respectively compared with Exp 6 and Exp 5. It may be that the affix and bi-gram phrase features carry some overlapping information, and contribute to some conflict. We assumed that more bi-gram phrase features can make more contribution to the classification, and performed experiments including low frequency bi-gram phrases (Exp 8), and the results shows some noise was introduced into classification with a slight F-value drop by 0.005.

6.2 Effect of post processing

The results of post-processing are reported in Table 7. Using the best model in the ME classification (Exp 6) as the baseline we applied tag changes and boundary correction to it. By using method 1 to fix the invalid I tags, the precision is degraded by 0.055, but with a moderate increase in recall (0.010) and decrease in F-score by 0.022 (Exp 10). Error analysis shows many false positives such as single lower case words. After correcting for invalid tag sequences (Exp 11) there is a slight increase in F-score (0.012). Next we used the Experiment 3 results as the second baseline on which to apply the boundary corrections.

The right boundary correction (Exp 12) further increases in F-score by 0.012 and the left boundary correction (Exp 13) also has a positive effect of 0.004. Applying both left and right boundary corrections there is a total increase of 0.015 in F-score. The left boundary correction only gives a slight positive effect, suggesting that the left boundary is more difficult to detect than the right boundary. The results of Experiment 15 show that the reclassification according to the head nouns has a positive effect on performance, improving the overall F-value by 0.044 compared to experiment 14.

The combined effect of post-processing is very effective, improving the performance over the ME model baseline by 0.071.

6.3 The Results

Table 8 shows the precision, recall and F-scores of the most populous categories of NE. The Protein class has the highest values for precision and recall. This is possibly due to proteins being the most frequent entity category in the training set. The Othername class is the second most frequent category, but it does not have a comparably high F-score. This is possibly due to the fact that Othername consists of some nested named entities which cause overlapping between Othername and other categories. Some small categories have very low F-score, due to a lack of training data.

Category	P	R	F	Freq
Protein	0.739	0.743	0.741	33.07%
OtherName	0.653	0.643	0.648	25.43%
DNA	0.718	0.642	0.678	11.69%
CellType	0.758	0.714	0.735	8.09%
CellLine	0.696	0.640	0.667	5.06%
Lipid	0.654	0.464	0.543	2.26%
Overall	0.700	0.666	0.682	100%

Table 8. Performance of major entity categories.

Exp. #	Processing	P	R	F	Effect	Baseline
9	**Baseline**	0.626	0.596	0.611	-	-
10	**Change invalid I to B**	0.571	0.606	0.588	-0.022	9
11	**Fix invalid tag sequence**	0.639	0.608	0.623	0.012	9
12	**Right Boundary Correction**	0.651	0.619	0.635	0.012	11
13	**Left Boundary Correction**	0.643	0.612	0.627	0.004	11
14	**Boundary Correction on both Side**	0.655	0.623	0.638	0.015	11
15	**Fix tag sequence according to head nouns**	**0.700**	**0.666**	**0.682**	**0.044**	**14**

Table 7. Effect of post-processing

The partial matching performance and identification performance are presented in Table 9 with the performance of exact match, left boundary correct, right boundary correct and identification only.

Boundary Performance	P	R	F
Exact match	0.700	0.666	0.682
Left Boundary	0.722	0.687	0.704
Right Boundary	0.739	0.703	0.721
Identification Only	0.802	0.762	0.781

Table 9. Partial matching and identification

The left boundary and right boundary have a higher performance than exact match, by .022 and .039 respectively in F-score. The results also show that the right boundary identification is better than left boundary identification. This shows that the left boundary is more difficult to detect, probably because of the difficulty in determining whether a modifier should be included in an NE or not. Identification outperforms classification by 0.099 F-value.

	P	R	F
Shen et al.[9]	0.677	0.653	0.665
Lee et al.[6]	0.718	0.698	0.708
Zhou et al.[17]	0.727	0.698	0.712
Lin et al.[8]	0.727	0.715	0.721
Experimental System	0.700	0.666	0.682

Table 10. A comparison to other systems

In Table 10 we show a comparison of our results to other systems. Although the test data is not exactly the same in each system, but for a rough comparison, our system achieved a performance close to these systems, and our system outperformed Shen's system slightly. Our system reported a relatively high boundary identification results, we think this is because the unigram entity feature and bi-gram phrase feature contributed to improve boundary identification.

6.4 Error analysis

Large numbers of misclassifications arise between the DNA and Protein classes. In the total of misclassified words, about 75% of the incorrectly recognized DNA terms are classified as Protein. This is due to the high overlap between these two classes. Another two categories that cause confusion is the CellLine of which 72% are incorrectly classified as CellType. These kinds of problems will most probably be addressed by exploring more contextual information.

Recognition error arises in some hyphen suffixes. For example, the entity "AP-1 –binding activity" of Othername has been partially recognized as AP-1 $_{\text{Protein}}$ –binding $_{\text{Outside}}$ activity $_{\text{Othername}}$. Similar situations are confronted with some high frequency hyphen suffixes, such as the word "cell-specific". This problem may be solved by a more careful study of hyphenated word features.

Abbreviation is another source of misclassification. The orthographic feature cannot capture enough information on abbreviations, because most abbreviations share the same orthographic feature. For example, the name "LPL" of Protein has always been recognized as Lipid.

True negatives are almost always identified by the feature of LowcaseOnly but are confounded with some entities. For example, the phrase "protein products" is never correctly labelled by our recognizer. These errors might be detected by using a dictionary, or exploration of more context information.

Other sources of errors are a number of non-entity words that are common medical terms classified as entities. Some high frequency words, such as stop words are incorrectly classified as Othername, as they sometimes appear in the composition of long entity names, for example, the word "family" and the word "and" have often been recognized as NE.

7 Conclusion and future work

In this paper we have presented a machine learning system for recognizing entity classes in biomedical abstracts. We have studied various linguistic features such as orthography, part of speech, affixes, head nouns, unigram terms and bigram phrases. We have also used simple rule based methods to correct invalid tag sequences and entity boundary errors.

We have achieved close to state of the art performance using very simple rule based post-processing without exploiting dictionaries. Our system achieved relatively high performance on boundary detection. However, there is still a 10% F-score gap between the identification performance and classification performance. This suggests that we have the potential to achieve better performance by looking at more informative features for semantic classification. In future work we will pursue better definitions of phrase forming rules and separate out the predictive value of different features for different entity types which is clearly shown to be operating in the use of the orthographic feature.

References

[1]. N. Collier, C. Nobata, and J. Tsujii. *Extracting the names of genes and gene products with a hidden Markov model.* In Proceedings of *COLING 2000*, pp 201-207, 2000.

[2]. N. Collier, K. Takeuchi. *Comparison of character-level and part of speech features for name recognition in biomedical texts.* J Biom. Inform. 37. pp423-435. 2004.

[3]. K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. 1998. *Toward information extraction: identifying protein names from biological papers. In Proc. of the Pacific Symposium on Biocomputing'98 (PSB'98)*, pp 707-718..

[4]. R. Gaizauskas, G. Demetriou and K. Humphreys. *Term Recognition and Classification in Biological Science Journal Articles. 2000.* In Proc. of *the Computional Terminology for Medical and Biological Applications Workshop* pp 37-44. 2000

[5]. L. Hirschman, A.A. Morgan, and A.S. Yeh, *Rutabaga by any other name: extracting biological names. J Biomed Inform*, 2002. 35(4): p. 247-59.

[6]. J. Kazama, T. Makino, Y. Ohta, J. Tsujii. *Tuning Support Vector Machines for Biomedical Named Entity Recognition.* In: Proceedings of *Workshop on NLP in the Biomedical Domain, ACL 2002.* pp1-8. 2002.

[7]. K.-J. Lee, Y.-S. Hwang, and H.-C. Rim. *Two-phase biomedical NER recognition based on SVMs*. In *Proceedings of ACL 2003, 2003.*

[8]. Y. Lin, T. Tsai, W. Chou, K. Wu, T. Sung and W. Hsu: *A Maximum Entropy Approach to Biomedical Named Entity Recognition*, In: Proceeding of *the 4th Workshop on Data Mining in Bioinformatics*: pp 56-61, 2004

[9]. D. Shen, J. Zhang, G. Zhou, S. Jian and L. Tan, *Effective Adaptation of a Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain*, In: Proceedings of ACL *2003 Workshop on NLP in Biomedicine, Sapporo, Japan*, pp49-56, 2003.

[10]. B. Settles. *Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets*. In: Proceedings of *the COLING 2004 NLPBA,*. 2004, pp 104-108, 2004.

[11]. B. Sundheim *Overview of the results of the MUC-6 evaluation*. In: Proceedings of *the Sixth Message Understanding Conference. Los Altos, CA: Morgan Kaufman; 1995*. p. 13-31

[12]. K. Takeuchi, and N. Collier. *Bio-medical Entity Extraction using Support Vector Machines.* In: Proceedings of *NLP in Biomedicine, ACL 2003. Sapporo, Japan*, pp 57-64, 2003.

[13]. Tsai, T.-H., Wu, S.-H., & Hsu, W.-L. (2005). *Exploitation of linguistic features using a CRF-based biomedical named entity recognizer.* to appear in ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, Detroit

[14]. Y. Tsuruoka, Y. Tateishi, . Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, *Developing a Robust Part-of-Speech Tagger for Biomedical Text*, Proceedings of *the 10th Panhellenic Conference on Informatics, 2005*

[15]. Tuason, O., L. Chen, H. Liu, J.A. Blake, and C. Friedman. *Biological Nomenclature: A Source of Lexical Knowledge and Ambiguity*. In: Proceedings of *Pac Symp Biocomput*. 2004. p. 238-49.

[16]. G. Zhou, *Recognizing Names in Biomedical Texts using Hidden Markov Model and SVM plus Sigmoid*, In: In: Proceedings of *the COLING 2004 NLPBA, Geneva, Switzerland*. 2004.

[17]. G. Zhou and J. Su. *Named Entity Recognition using an HMM-based Chunk Tagger*. In Proc. of *the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 473-48*0 2002.

[18]. C. Nobata, N. Collier and J. Tsujii. *Automatic term identification and classification in biology texts*. In Proc. of *the 5th NLPRS, pp 369-374.* 1999.

[19]. M. Krauthammer, Rzhetsky A, Morozov P, Friedman C. *Using BLAST for identifying gene and protein names* in journal articles. Gene 2000;259(1–2):245–52. 2000

[20]. J. Kim, T. Ohta, Y. Teteisi, and J. Tsujii. *GENIA corpus – a semantically annotated corpus for bio-textmining. Bioinformatics 19 (suppl.1)* 2003