Boosting Performance of Bio-Entity Recognition by Combining Results from Multiple Systems

Luo Si
Language Technology Inst
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
lsi@cs.cmu.edu

Tapas Kanungo
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
kanungo@us.ibm.com

Xiangji Huang
School of Information Technology
York University
Toronto, Canada
jhuang@yorku.ca

ABSTRACT

The task of biomedical named-entity recognition is to identify technical terms in the domain of biology that are of special interest to domain experts. While numerous algorithms have been proposed for this task, biomedical named-entity recognition remains a challenging task and an active area of research, as there is still a large accuracy gap between the best algorithms for biomedical named-entity recognition and those for general newswire named-entity recognition. The reason for such discrepancy in accuracy results is generally attributed to inadequate feature representations of individual entity recognition systems and external domain knowledge.

In order to take advantage of the rich feature representations and external domain knowledge used by different systems, we propose several Meta biomedical named-entity recognition algorithms that combine recognition results of various recognition systems. The proposed algorithms – majority vote, unstructured exponential model and conditional random field – were tested on the GENIA biomedical corpus. Empirical results show that the F score can be improved from 0.72, which is attained by the best individual system, to 0.96 by our Meta entity recognition approach.

Categories & Subject Descriptors:

H.3.3 [Information Search and Retrieval]: Text Mining

General Terms:

Algorithms

Keywords:

Biomedical named-entity recognition; Meta recognition

1. INTRODUCTION

Biomedical literature contains a rich set of biomedical entities and information regarding the relationships and interactions among these entities. These entities and their relationships are especially useful for biologists in their quest for information [11]. The exponential growth of available biomedical literature on the Web and publicly accessible databases requires intelligent information systems that help researchers to search and analyze information. Therefore, the use of computational techniques to automatically extract useful information from biomedical texts has received increasing attention. Furthermore, to perform higher level biomedical information extraction tasks such as event extraction, summarization and question answering, most systems first identify technical terms in the domain of molecular biology that are of special interests to domain experts [11]. This is called named-entity recognition in natural language processing community [6].

The named-entity recognition task for general-purpose domain such as newswire data has been studied for a long time [3,6,22]. Both handcrafted linguistic rule based methods and machine learning based methods have been proposed for this task. Machine learning based methods [3,6] have attracted particular interest as they avoid the laborious task of manually deriving linguistic rules, and also because they can be easily adapted to new domains and new languages. Good progress has been made in named-entity recognition of newswire data and best algorithms can now achieve 'near human' performance (e.g., F score of about 0.95) [3,6,22].

The named-entity recognition task in the biomedical domain has different characteristics from that in the newswire domain. Authors tend to use more diverse notations for biomedical entities. In addition, biomedical named-entities usually have much more diverse capitalization patterns than those in newswire domain. A richer set of features, therefore, should be used to represent biomedical entities [11].

A large body of machine learning algorithms has been proposed for biomedical named-entity recognition such as hidden Markov model (HMM) [8,17,19,24,25], support vector machine (SVM) [4,13,16,19,23,25], maximum entropy markov model (MEMM) [7,14] and conditional random field (CRF) [12,15,20,23]. In order to capture the diverse characteristics of biomedical entities, different sets of features such as lexical features, affix information, orthographic features or even external resources such as gazetteers [7,25] or WWW [7,20] have been incorporated into different algorithms.

BIOKDD'05, August, 2005, Chicago, Illinois, USA.
Copyright 2005 ACM 1-xxxxx-xxx-xxxxxxxxx…$5.00.

However, biomedical named-entity recognition still remains a challenging problem [11]. Despite the near-perfect performance of named-entity recognition in newswire data, similar methods do not work so well in biomedical domain and there is a large accuracy gap of about 20 points in the F score [6,9,11,25]. This problem suggests that individual biomedical named-entity systems may not cover entity representations with enough rich features and no single type of algorithm is optimal to achieve the best performance.

One natural idea of boosting performance of biomedical named-entity recognition is to combine the results of multiple biomedical entity recognition systems. This approach provides us the opportunity to combine results from multiple systems that collectively use rich and diverse feature representations and also take the advantage of utilizing multiple algorithms for achieving higher recognition accuracy.

Similar approach of combining results from multiple systems has been successfully applied in information retrieval community [1], where retrieved ranked lists from multiple information retrieval systems are combined together into a final ranked list. Empirical evidence has demonstrated that Meta retrieval approach substantially improves retrieval accuracy. However, Meta retrieval method is different from Meta entity recognition method as Meta retrieval method combines unstructured results of ranked lists while Meta entity recognition combines structured results from different named-entity recognition systems.

In this paper we propose three methods for Meta biomedical named-entity recognition. The first method uses majority vote from a set of entity recognition systems to produce combined results. This simple method does not require any training data. The second method trains an unstructured exponential model and uses the recognition results from individual systems as features to predict the correct recognition result for each word in test sentence separately. Finally, a more sophisticated structured line chain conditional random field model [12] is applied. This model utilizes structure information regarding transition among different types of entities. Although some of these techniques have been applied in other applications, to our knowledge they have never been used for Meta biomedical entity recognition.

An extensive set of empirical study has been conducted on the GENIA [1] corpus [10,11] with the task of identifying five different types of biomedical named-entities. Entity recognition results from eight different systems are considered in the Meta recognition system for combination. The best single system achieves an F score of 0.72 on the GENIA corpus [25], while the Meta recognition system with the linear chain conditional random field model achieves an F score of about 0.96. This large improvement demonstrates the power of combining multiple results for the biomedical named-entity recognition task. Furthermore, a careful comparison among different Meta recognition algorithms shows that the supervised methods of unstructured exponential model and linear conditional random field method are more effective than the simple majority vote algorithm. The structured conditional random field model achieves higher accuracy than the unstructured exponential model, which demonstrates the advantage of utilizing structure information among named-entity recognition results.

In the next section we discuss prior research related to biomedical name-entity recognition algorithms and Meta retrieval technology in information retrieval. In Section 3 we describe the three proposed Meta entity recognition algorithms --- majority vote, unstructured exponential model and structured conditional random field model. We outline the experimental methodology in Section 4 and finally present the results of our empirical study in Section 5. In Section 6 we conclude by summarizing our work and pointing out a few future research directions.

2. RELATED WORK

The approach proposed in this paper combines results from multiple biomedical named-entity recognition systems. In the next subsection we discuss specific algorithms for Bio-Entity recognition, and in the subsequent subsection we describe Meta retrieval algorithms used in information retrieval.

2.1 Algorithms for Bio-Entity Recognition

Biomedical named-entity recognition is still an active research topic, and numerous algorithms have been proposed using different feature representations. For example, in the JNLPBA [10,11] shared task of Bio-entity recognition task, eight entity recognition systems utilize different learning algorithms and different sets of features. The algorithms include variants of Support Vector Machine (SVM) [4,13,16,19,23,25], Hidden Markov Model (HMM) [8,17,19,24,25], Maximum Entropy Markov Model (MEMM) [7,14] and Conditional Random Field (CRF) Model [12,15,20,23].

Besides learning algorithms, feature representation has been recognized as a crucial factor to get good performance in Bio-Entity recognition. In the JNLPBA task [10,11], lexical features are widely used among many systems as biomedical named-entities generally have a different vocabulary from general English words. When SVM-based systems have trouble to incorporate large size of lexical features, different generalization of lexical features such as prefixes or suffixes (e.g., suffixes as ~in or ~ase for protein names) are utilized. Furthermore, some general features such as part of speech tags or word shapes as well as domain specific features such as gene sequences are also utilized in different systems. More detail can found in [11].

In addition to using features from the biomedical document itself, many systems tend to use gazetteers and other external resources for better generalization performance. Some systems use gene names from biomedical websites such as LocusLink [7] or Gene Ontology [7,13], while some other systems use the Web and construct lexicon [19,20] by collecting words that frequently appear in context with known biomedical named-entities.

To summarize, a large body of learning algorithms is available for biomedical named-entity recognition. They utilize diverse feature representations. It can be expected that the recognition results from these systems are also diverse and complementary. In the light of these facts, we believe that a good Meta biomedical named-entity recognition algorithm can take

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

advantage of the diversity of the results from multiple systems and improve the results further.

2.2 Meta Retrieval Algorithm

The approach of combining results from multiple systems has been successfully utilized in the information retrieval community [1,5,18].

Simple methods like Borda Count [1] do not require training data and favor documents that are retrieved by more individual systems against documents that are retrieved by fewer or no systems. More sophisticated algorithms that utilize training data include Naive Bayesian method [1] and logistic regression model [5]. The Naive Bayesian method makes an independence assumption among results from multiple systems, which may be inaccurate in many cases. The logistic regression model does not make the independence assumption and uses retrieved results from multiple systems as features to predict the probability of relevance for each document candidate. It has been shown that this method achieves satisfactory Meta combination results.

Although some Meta retrieval algorithms have been proposed for information retrieval, they cannot be directly used for the Meta biomedical named-entity recognition task. In particular, Meta retrieval algorithms treat only the binary case -- relevance or irrelevance of any retrieved document -- while biomedical named-entity recognition generally involves multiple types of named-entities. In addition, information retrieval systems provide unstructured ranked lists while name-entity recognition systems provide structured results of annotated sentences. These characteristics of Meta biomedical named-entity recognition task are investigated in the next section in detail.

3. ALGORITHMS FOR META BIO-ENTITY RECOGNITION

In this section, we present three algorithms for Meta biomedical named-entity recognition. All the three algorithms deal with recognition of multiple types of biomedical named-entities. The first algorithm is a simple majority vote algorithm that requires no training; the second is an unstructured exponential model that learns relative weights but does not incorporate structure information, and the third is a conditional random field model that takes full advantage of the structure information among biomedical named-entities and learns relative weights.

We now introduce the formal notation used in this paper. Let an annotated sentence be composed of words of $\vec{w}_i$ and annotated entities $\vec{s}_i$. The training data is comprised of I annotated sentences: $D = \{(\vec{w}_1, \vec{s}_1), (\vec{w}_2, \vec{s}_2)_{\ldots\ldots,} (\vec{w}_I, \vec{s}_I)\}$, where the pair $(\vec{w}_i, \vec{s}_i)$ denotes the ith annotated sentence. We assume that the ith annotated sentence contains N_i words and denote the jth surface word and the corresponding named-entity by (w_{ij}, s_{ij}). We associate a category value for each type of named-entity and an additional "Non-entity" category for general English words. Each s_{ij} can attain any of the K category values. Assume that L annotated results are provided from L biomedical named-entity recognition systems. Thus, for the ith sentence the lth system's candidate results are denoted as: $\{c_{l\_i1}, c_{l\_i1}, \ldots\ldots, c_{l\_iN_i}\}$, where each item has a category value out of K choices. Finally, the task of Meta named-entity recognition algorithm is to combine the L candidate entity recognition results into a single result $\vec{s}_t$ for each test sentence t.

3.1 Simple Majority Vote Algorithm

The majority vote algorithm assumes that named-entities are correctly recognized by most individual systems, while different systems make mistakes at different places [1].

Let us introduce the binary indicator feature function $f(k, c_{l\_tj})$, which has a value 1 when the lth entity recognition system annotates the jth word in the test sentence as the entity of type k, and 0 when this is not true. Then the recognition rule of majority vote algorithm can be described formally as follows:

$$\hat{S}_{tj} = \arg\max_{k} \sum_{l} f(k, c_{l\_tj}) \qquad (1)$$

where t represents the test sentence and $\hat{S}_{tj}$ is the annotated entity result for the jth word in the test sentence.

One particular issue about majority vote is that votes from inaccurate entity recognition systems may not be reliable and may deteriorate the final results. Therefore, a variant of majority vote algorithm, which only considers votes from top few accurate systems, is often used in practice. This algorithm is also considered in this paper.

3.2 Unstructured Exponential Model Algorithm

One problem with the majority vote algorithm is that it treats the votes from different entity recognition systems equally. However, it is clear that more accurate systems should have more influence for the final decision than less accurate systems. The unstructured exponential model algorithm automatically derives appropriate weights for different systems from the training data, which means that those systems that are more accurate on training data are assigned with larger weights to recognize entities on test data. This type of bias is reasonable as long as the training data is representative.

Formally, the lth individual biomedical named-entity recognition system is associated with a weight λ_l and the probability of assigning entity of category k to the jth word in ith sentence is calculated as:

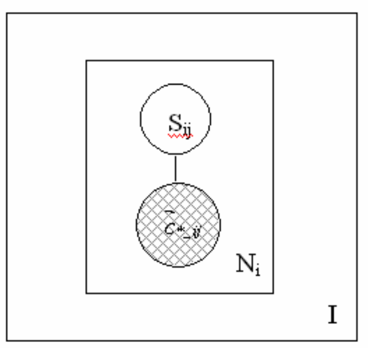


Figure 1. Graphical representation of unstructured exponential model (shared part is observed $\vec{c}_{*\_ij}$ as features from multiple entity recognition systems). Given the entity candidate features from multiple systems and model parameters, the named-entities are generated for each word separately.

$$P(\hat{S}_{ij}=k\,|\,\{w_{ij},c_{*\_ij}\})=\frac{\exp(\sum_{l}\lambda_l f(k,c_{l\_ij}))}{\sum_{k'}\exp(\sum_{l}\lambda_l f(k',c_{l\_ij}))} \qquad (2)$$

Note that no feature from the surface word itself is used in the current formulation yet. It may be useful to incorporate surface word features for more complicated combination strategy. However, empirical study in Section 5 demonstrates that this model can achieve very good performance with very limited amount of training data. Adding a lot of surface word features may cause overfitting problem with limited amount of data.

In fact, the exponential model can be seen as a multi-category extension of the logistic regression model for Meta retrieval system of information retrieval [1,5]. The graphical representation of this probabilistic model is shown in Figure 1. It can be seen from Figure 1 that given the entity features from multiple systems and model parameters, the named-entities are generated for each word separately without any interaction. That is why this model is called unstructured model.

The training criterion of this model is to maximize the conditional log-likelihood of the training data. Formally the parameter estimation problem is:

$$\vec{\lambda}^{*}=\arg\max_{\bar{\lambda}}\sum_{i,j,k}P(\hat{S}_{ij}=k)\log P(\hat{S}_{ij}=k\,|\,\{w_{ij},c_{*\_ij}\}) \qquad (3)$$

Where $P(\hat{S}_{ij}=k)$ is the empirical probability distribution for different types of named-entities of a specific word. It is 1 for one type of name-entity and zero for all the others.

The objective function in Equation (3) is a convex function and the optimization method of iterative scaling is used to obtain optimal parameter value. More detailed information about the iterative scaling method can be found in [2].

3.3 Conditional Random Field Algorithm

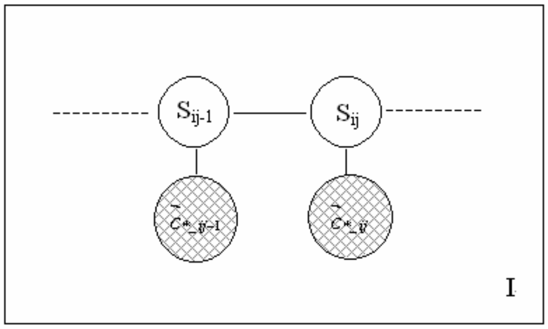


Figure 2. Graphical representation of linear conditional random field model (shadowed part is observed $\vec{c}_{*\_i*}$ as features from multiple entity recognition systems.) Given the entity candidate features from multiple systems and model parameters, the named-entities within a sentence are generated with interaction.

One important piece of useful information that is missing in the unstructured exponential model is the structure information. The named-entities assigned to nearby words are actually correlated with each other. If the previous word is recognized as a part of protein name, it is likely that the current word has a higher probability to be a part of protein entity than a cell_line entity. Conditional random field method [12] can be used to model the correlation between biomedical named-entities.

More specifically, the conditional random field model calculates conditional probabilities for whole annotated sentences instead of individual entities. In this paper, a linear chain conditional random field model is used. This is formally represented as:

$$P(\vec{s}_i=\{k_1,\dots,k_j,\dots,k_{N_i}\}\,|\,\vec{w}_i,c_{*\_i*}\}) = \frac{\exp(\sum_{j}\sum_{m}\lambda_m f_m(k_{j-1},k_{j,}c_{*\_i*}))}{\sum_{k_1',..k_j',...}\exp(\sum_{j}\sum_{m}\lambda_m f_m(k_{j-1}',k_j',c_{*\_i*}))} \qquad (4)$$

In particular, each feature function is associated with two concatenated entities and the corresponding candidate entity results from multiple entity recognition systems. The graphical model of linear chain conditional random field is shown in Figure 2. It can be seen that adjacent named-entities are associated with each other. This characteristic allows the conditional random field method to take advantage of structure information among entities.

The training criterion of conditional random field has a similar objective function to that of unstructured exponential model:

$$\vec{\lambda}^{*}=\arg\max_{\bar{\lambda}}\sum_{i}\log(P(\vec{s}_i\,|\,\{\vec{w}_i,c_{*\_i*}\})) \qquad (5)$$

The conditional likelihood function involves a sentence-scale normalization factor as indicated in Equation (4); the training computational complexity is much larger than that of unstructured exponential model. Quasi-Newton optimization method [21] has been shown to be more efficient than several other alternatives such as conjugate gradient and iterative scaling. This method is used in this work to train the linear chain

	Protein	DNA	RNA	Cell_type	Cell_line	All
Num of occurrences	5,067	1,056	118	1,921	500	8,662
Percent of total words	12.5%	2.6%	0.3%	4.8%	1.2%	21.4%

Table 1. Num of occurrences and percentage of total words for five types of biomedical named-entities in the corpus.

	Zho [25]	Fin [7]	Set [20]	Son [23]	Zha [24]	Rös [19]	Par [16]	Lee [13]
Recall	0.760	0.716	0.703	0.678	0.691	0.674	0.665	0.508
Precision	0.694	0.686	0.693	0.648	0.610	0.610	0.598	0.476
F-Score	0.726	0.701	0.698	0.663	0.648	0.640	0.630	0.491

Table 2. Performance of individual systems. Systems are ranked by their F scores from the highest (Left) to the lowest (Right).

conditional random field model for Meta biomedical named-entity recognition.

Given the estimated model, the recognition step of conditional random field is also more complicated than that of exponential model. A dynamic programming solution is utilized here to calculate the most likely named-entity sequence given the test sentence. Specially, a forward-backward inference algorithm like that for HMM is applied. The 'forward value' $a_j(S_{tj}=k)$ is defined as the probability of being in entity of type *k* at *j*th position given the observation up to time *j* and $\beta_j(S_{tj}=k)$ is the probability of being in entity of type *k* at *j*th position given the observation after time *j*. Recursive steps are applied to calculate the whole set of forward and backward values:

$$\begin{aligned} & a_{j+1}(S_{tj+1}=k) \\ & \quad = \sum_{k'} a_j(k')\exp(\sum_m \lambda_m f_m(k',k,c_{*\_tj+1})) \\ & \beta_j(S_{tj}=k) \\ & \quad = \sum_{k'} \exp(\sum_m \lambda_m f_m(k,k',c_{*\_tj+1}))\beta_{j+1}(k') \end{aligned} \qquad (6)$$

Viterbi algorithm is applied with forward and backward values and finally the optimal sequence of named-entities is computed.

4. EXPERIMENTAL METHODOLOGY

We used the entity recognition results from eight different biomedical named-entity recognition systems that participated in the JNLPBA competition [2]. In the JNLPBA competition [11], each entity recognition system is required to recognize five types of entities as protein, DNA, RNA, cell_type and cell_line within documents in the GENIA corpus [10]. We utilize these results to construct Meta biomedical entity recognition system in this paper.

The recognition results are evaluated using the F score. F score is defined as: $F=(2PR)/(P+R)$, where P denotes Precision, which is the ratio of the number of correctly recognized named-entities to the number of recognized named-entities. R denotes Recall, which is the ratio of the number of correctly recognized entities to the number of true entities [11].

Since the eight systems provide results only on the test set of JNLPBA task that contains 404 documents of the GENIA corpus, we split the test data of JNLPBA into training and test data for our experiments. There are altogether 404 Medline abstracts, which are composed of 4260 sentences. The biomedical entity distribution is tabulated in Table 1. In order to fully investigate the behavior of different Meta recognition algorithms, two different training configurations were used in this work: i) 10 annotated documents for training and ii) 5 annotated documents for training. The 5 (or 10) documents that contain all the five types of annotated biomedical named-entities were randomly chosen from the 404 abstracts as training data and the remaining documents were used as test data. The training set has about 50 (or 100) sentences with about 1,250 (or 2,500) words. The random split process was repeated five times for each experiment and the evaluation results were averaged.

The performance of eight different systems on the whole corpus (404 abstracts and no training) is shown in Table 2. Three out of eight systems achieve F score around 0.7 while the F-score of other systems ranges from 0.5 to 0.65.

5. EXPERIMENTAL RESULTS

In this section we present the results of applying the proposed Meta biomedical named-entity recognition algorithms on the GENIA corpus and compare these results to individual systems. Two particular issues are investigated by the empirical study in this section:

1. Whether Meta biomedical named-entity recognition approach improves recognition accuracy over individual systems, and how do different Meta biomedical entity recognition algorithms compare against each other?
2. Detailed analysis for different types of named-entities is provided to carefully compare the results from individual systems and different Meta recognition algorithms.

5.1 Overall Recognition Accuracy

The first set of experiments was conducted to study the effectiveness of the simple majority vote algorithm. In order to show the full spectrum of its behavior, we vary the number of systems that are considered for voting. In particularly, we sort all the systems by their F scores as shown in Table 2 and use the simple majority vote algorithm to combine the results from best

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

	B1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
Recall	0.761	0.876	0.859	0.850	0.786	0.797	0.770	0.778
Precision	0.696	0.739	0.802	0.771	0.724	0.727	0.712	0.707
F-Score	0.727	0.802	0.830	0.808	0.754	0.761	0.740	0.741

Table 3. Performance (in F score) of simple majority vote algorithms compared with the best single system (10 documents are used for training and results are averaged by five random splits). Simple majority vote algorithms combine results from different number of top systems (B1: best single system; M_2 means combination of two most accurate systems and so on).

	B1 (Baseline)	**M_8**		**M_3**		**EXP**			**CRF**		
		F Score	**Impr(%)**	**F Score**	**Impr(%)**	**F Score**	**Std**	**Impr(%)**	**F Score**	**Std**	**Impr(%)**
Recall	0.761	0.778	(+2.2%)	0.859	(+12.9%)	0.926	0.016	(+21.7%)	**0.956**	0.012	**(+25.6%)**
Precision	0.696	0.707	(+1.6%)	0.802	(+15.2%)	0.920	0.021	(+32.2%)	**0.971**	0.010	**(+39.5%)**
F-Score	0.727	0.741	(+1.9%)	0.830	(+14.2%)	0.923	0.015	(+27.0%)	**0.964**	0.011	**(+32.6%)**

Table 4. Performance of Meta biomedical named-entity systems compared with the best single system (10 documents are used for training and results are averaged by five random splits; F Score: F measure; Std: standard deviation across 5 random splits; Impr(%): Relative improvement over baseline). B1: Best single system; M_8: majority vote from eight systems; M_3: majority vote from best three systems; EXP: unstructured exponential model: CRF: conditional random field. (Standard deviation of M_8 and M_3 are not reported as they are very small)

	B1 (Baseline)	**M_8**		**M_3**		**EXP**			**CRF**		
		F Score	**Impr(%)**	**F Score**	**Impr(%)**	**F Score**	**Std**	**Impr(%)**	**F Score**	**Std**	**Impr(%)**
Recall	0.759	0.777	(+2.4%)	0.858	(+13.0%)	0.907	0.026	(+19.4%)	**0.921**	0.024	**(+21.3%)**
Precision	0.694	0.706	(+1.7%)	0.801	(+15.4%)	0.879	0.035	(+26.7%)	**0.953**	0.018	**(+37.3%)**
F-Score	0.725	0.740	(+2.0%)	0.829	(+14.3%)	0.893	0.030	(+23.3%)	**0.937**	0.021	**(+29.2%)**

Table 5. Performance of Meta biomedical named-entity systems compared with the best single system (5 documents are used for training and results are averaged by five random splits; F Score: F measure; Std: standard deviation across 5 random splits; Impr(%): Percentage improvement over baseline). Algorithm descriptions are the same as the above.

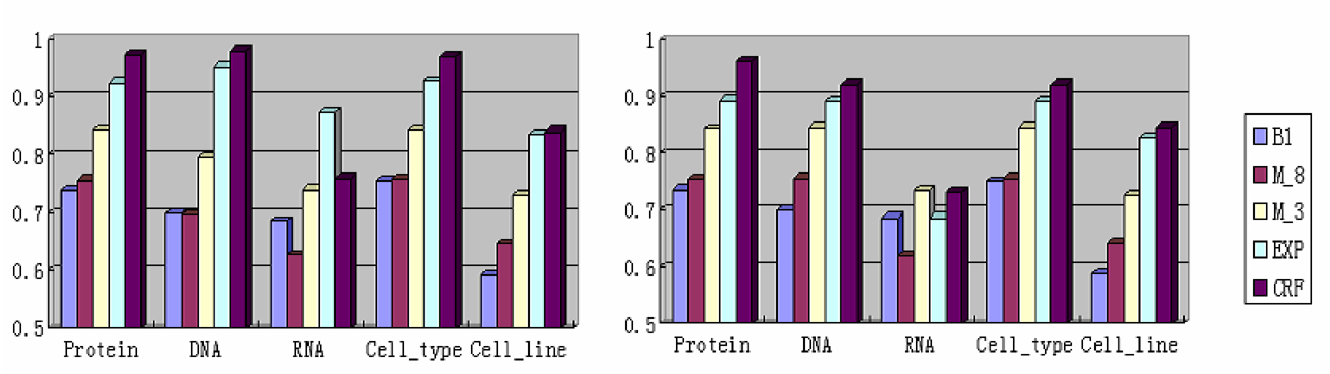
two systems (M_2), best three systems (M_3) and so on. The detailed experiments are shown in Table 3. While the majority vote algorithm does not have to be trained, we made the experimental setup identical to that used for the trainable Meta algorithms to make the evaluation results comparable: 10 documents were held for training in each of the five random splits and the remaining 394 documents were used for test (the results when 5 documents were used for training are almost identical with these results and are not shown). The majority voting algorithms did not use the 10 (and 5) training documents – only the trainable algorithm made use of them.

Note a particular issue of simple majority vote algorithm is tie breaking. If the votes from multiple systems are the same for some entities, the preference is given in the order to protein, DNA, RNA, cell_ type, cell_line and "Non-entity".

It can be seen from Table 3 that simple majority vote algorithm does achieve more accurate result than single best system. However, its performance varies with the number of systems of combination. The best results are achieved when top three or four systems are considered for voting and the accuracy drops significantly when more and more low accuracy systems are added into the combination. This behavior suggests that appropriate weights should be assigned to individual systems in order to achieve optimal performance of Meta named-entity recognition; and this is exactly the goal of the unstructured exponential model and conditional random field model

More experiments were conducted to study four types of Meta biomedical entity recognition algorithms. The algorithms are: M_8 (majority vote algorithm form all of the eight individual systems); M_3 (majority vote algorithm from three most accurate individual systems as Zho [25], Fin [7] and Set [20]); EXP (unstructured exponential model) and CRF (conditional random field model). Both the EXP and CRF algorithms take advantage of training data. Table 4 shows the results when 10 documents were available for training. It can be seen that EXP and CRF achieve a significant improvement over the best single system and also are much more accurate than the simple majority algorithm. More careful analysis shows that EXP and CRF algorithms automatically assign appropriate weights for individual systems. For example, EXP assigns more weights to the top three systems than the other systems. Furthermore, CRF algorithm generates more accurate results than the EXP algorithm. This demonstrates the power of utilizing the structure information among entities.

Another set of experiments was designed to test the behavior of different Meta entity recognition algorithms with more limited amount of training data. The experiments shown in Table 5 use only 5 documents as training data. It can be seen from Table 5 that the performance of M_8 and M_3 algorithms remain at



(a)

(b)

Figure 3. Performance of best single systems and Meta recognition algorithms for different types of biomedical named entities. (a) is the case with 10 documents for training while (b) is the case with 5 documents for training. For B1, system from Zho [21] is used to predict protein, DNA, cell_type and cell_line while the system from Fin [5] is used to predict RNA.

about the same level as those in Table 4 since these algorithms do not utilize training data and their accuracy does not depend on the size of training data. The accuracy of EXP and CRF algorithms drops slightly with more limited amount of training data. However, their advantage over best single system or simple majority vote recognition algorithm is still very large. This set of experiments suggests that Meta biomedical named-entity recognition algorithms can acquire very accurate results even with very limited amount of training data (i.e., about 50 training sentences).

Other configurations with more training data have also been studied. When 15, 20 or more documents are used for training, the accuracy of EXP and CRF methods increase. However, the improvement over the results of less training data (i.e., 5 or 10 documents.) is small due to the high performance of EXP and CRF methods with limited amount of training data.

Both unstructured exponential algorithm and conditional random algorithm are very efficient. They are implemented using Matlab. It takes about 30 seconds to train the exponential model and about 2 minutes to train the conditional random field model in the case of 10 training documents. It only takes about 30 seconds for CRF to generate combined results for 394 documents while several seconds for the exponential model.

5.2 Recognition Accuracy for Different Types of Biomedical Named Entities

This set of experiments shows how Meta entity recognition algorithms improve the recognition accuracy for each type of biomedical named entity.

Figure 3 shows the performance of best single system and Meta recognition algorithms for different types of biomedical named-entities. Note that different individual systems may be optimal for different types of biomedical named-entities. For example, the system by Fin [7] has a better performance for RNA entities than the system by Zho [25]. More detail can be found in [11].

It can be seen from Figure 3 that Meta recognition algorithms CRF, EXP and M_3 achieve better performance than single best system. Unstructured exponential model and conditional random field model achieve better result than other algorithms in most cases by assigning appropriate weights to the results from multiple systems. In fact, the weights of different systems are also varied for the recognition of different types of entities. Furthermore, the CRF method provides the most accurate results in most cases, which again demonstrates the power of utilizing structure information.

6. CONCLUSION AND FUTURE WORK

Due to the large vocabulary and very diverse notations of biomedical entities, the performance of current biomedical named-entity recognition systems is still not satisfactory. Possible reasons are inadequate feature representations of individual systems and ineffectiveness of individual algorithms.

This paper proposes a Meta biomedical named-entity recognition approach by combining results from multiple systems. Three types of Meta recognition algorithms are proposed. Empirical study shows that Meta biomedical named-entity methods can substantially improve recognition accuracy over individual systems. The best results are obtained with a conditional random field method that takes the advantage of structure information for recognition. With a small amount of training data, this method provides recognition results with an F score of 0.96 while the F score of the best single system is only 0.72 [11,25]

As more and more trainable biomedical named-entity systems are available, we will apply the Meta entity recognition approach on other biomedical corpus for more complete evaluation. Training data can be used to train both individual named-entity recognition systems and the Meta recognition system. Furthermore, more sophisticated model which considers surface word features to combine results will be investigated in future work.

REFERENCES

[1] J. A. Aslam and M. Montague (2001). Models for Metasearch. In *Proceedings of the 24th Annual*

International ACM SIGIR Conference on Research and Development in Information Retrieval.

[2] A. Berger. (1997). A gentle introduction to iterative scaling. http://www-2.cs.cmu.edu/~aberger/maxent.html

[3] D. M. Bikel, R. L. Schwartz and R. M. Weischedel. (1999). An algorithm that learns what's in a name. *Machine Learning*, vol. 34, no. 1-3, pp. 211-231, 1999.

[4] Christopher J.C. Burges. (1998) A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery* 2, 121-167.

[5] A. Le Calvé, J. Savoy (2000): Database Merging Strategy Based on Logistic Regression. *Information Processing & Management,* 36(3), 341-359.

[6] DARPA. (1995). *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD, USA, November. Morgan Kaufmann.

[7] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, G. Sinclair and C. Manning. (2004). Exploiting Context for Biomedical Entity Recognition: From Syntax to the Web. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, Geneva, Switzerland.

[8] T. Kanungo. HMM software learning toolkit. University of Maryland Institute for Advanced Computer Studies, http://www.cfar.umd.edu/~kanungo/software/software.html

[9] J. D. Kim, T. Ohta, Y. Tateisi and J. Tsujii. (2002). Corpus-Based Approach to Biological Entity Recognition. In *Proceedings of the Second Meeting of the Special Interest Group on Test Data Mining of ISMB (BioLink-2002)*, Edmonton, Canada.

[10] J. D. Kim, T Ohta, Y. Tateishi and J. Tsujii. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 (Suppl.1): 180-182.

[11] J. D. Kim, T Ohta, Y. Tateishi and J. Tsujii. (2004). Introduction to the Bio-Entity Recognition Task at JNLPBA. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, Geneva, Switzerland.

[12] J. Lafferty, A. McCallum and F. Pereira. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*. Williamstown, MA, U.S.A.

[13] C. Lee, W. J. Hou and H.-H. Chen. (2004). Annotating Multiple Types of Biomedical Entities: A Single Word Classification Approach. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, Geneva, Switzerland.

[14] A. McCallum, Dayne Freitag and Fernando Pereira. Maximum Entropy Markov Models for Information Extraction and Segmentation. (2000). In *Proceedings of the International Conference on Machine Learning*. Williamstown, MA, U.S.A.

[15] A. McCallum and W. Li. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Conference on Natural Language Learning*, pages 188–191. Edmonton, Canada.

[16] K. M. Park, S. H. Kim, D. G. Lee and H. C. Rim. (2004). Boosting Lexical Knowledge for Biomedical Named Entity Recognition. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004),* Geneva, Switzerland.

[17] L. R. Rabiner. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77:257—286.

[18] C. J. van Rijsbergen. (1979). Information Retrieval. Butterworths, London.

[19] M. Rössler. (2004). Adapting a NER-System for German to the Biomedical Domain. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004),* Geneva, Switzerland.

[20] B. Settles. (2004). Biomedical Named Entity Recognition Using Conditional Random Fields and Novel Feature Sets. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004),* Geneva, Switzerland.

[21] F. Sha and F. (2003). Shallow Parsing with Conditional Random Fields. In *Proceedings of Human Language Technology-NAACL 2003,* Edmonton, Canada.

[22] E. F. Tjong, K. Sang and F. De Meulder. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-2003)*, pages 142-147. Edmonton, Canada.

[23] Y. Song, E. Kim, G. Geunbae Lee and B. K. Yi. (2004). POSBIOTM-NER in the shared task of BioNLP/NLPBA 2004. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004),* Geneva, Switzerland.

[24] S. J. Zhao. (2004). Name Entity Recognition in Biomedical Text using a HMM model In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004),* Geneva, Switzerland.

[25] G. D. Zhou and J. Su. (2004). 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),* Geneva, Switzerland.