Introduction to the Bio-Entity Recognition Task at JNLPBA

Jin-Dong KIM, Tomoko OHTA, Yoshimasa TSURUOKA, Yuka TATEISI
CREST, Japan Science and Technology Agency, and
Department of Computer Science, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

Nigel COLLIER
National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan†

Abstract

We describe here the JNLPBA shared task of bio-entity recognition using an extended version of the GENIA version 3 named entity corpus of MEDLINE abstracts. We provide background information on the task and present a general discussion of the approaches taken by participating systems.

1 Introduction

Bio-entity recognition aims to identify and classify technical terms in the domain of molecular biology that correspond to instances of concepts that are of interest to biologists. Examples of such entities include the names of proteins, genes and their locations of activity such as cells or organism names as shown in Figure 1.

Entity recognition is a core component technology in several higher level information access tasks such as information extraction (template filling), summarization and question answering. These tasks aim to help users find structure in unstructured text data and aid in finding relevant factual information. This is becoming increasingly important with the massive increase in reported results due to high throughput experimental methods.

Bio-entity recognition by computers remains a significantly challenging task. Despite good progress in newswire entity recognition (e.g. (MUC, 1995; Tjong Kim Sang and De Meulder, 2003)) that has led to 'near human' levels of performance, measured in the high 90s for F-score (van Rijsbergen, 1979), similar methods have not performed so well in the bio-domain leaving an accuracy gap of some 30 points of F-score. Challenges occur for example due to ambiguity in the left boundary of entities caused by descriptive naming, shortened forms due to abbreviation and aliasing, the difficulty of creating consistently annotated human training data with a large number of classes, etc. In order to make progress it is becoming clear that several points need to be considered: (1) extension of feature sets beyond the lexical level (part of speech, orthography etc.) and use of higher-levels of linguistic knowledge such as dependency relations, (2) potential for re-use of external domain knowledge resources such as gazetteers and ontologies, (3) improved quality control methods for building annotation collections, (4) fine grained error analysis beyond the F-score statistics.

We have shown that <**cons sem="G#protein"**>*interleukin-1*</**cons**> (<**cons sem="G#protein"**>*IL-1*</**cons**>) *and* <**cons sem="G#protein"**>*IL-2*</**cons**> *control* <**cons sem="G#DNA"**>*IL-2 receptor alpha (IL-2R alpha) gene*</**cons**> *transcription in* <**cons sem="G#cell_line"**>*CD4-CD8- murine T lymphocyte precursors*</**cons**>.

Figure 1: Example MEDLINE sentence marked up in XML for molecular biology named-entities.

The JNLPBA shared task [1] is an *open challenge* task and as such we allowed participants to use whatever methodology and knowledge sources they liked in the bio-entity task. The systems were evaluated on a common benchmark data set using a common evaluation method. Although it is not directly possible to compare systems due to the diversity of resources used the F-score results provide an approximate indication of how useful each method is.

2 Data

The training data used in the task came from the GENIA version 3.02 corpus (Kim et al.,

* {jdkim,yucca,okap,tsuruoka}@is.s.u-tokyo.ac.jp
† collier@nii.ac.jp

[1] http://research.nii.ac.jp/~collier/workshops/JNLPBA04st.htm

2003). This was formed from a controlled search on MEDLINE using the MeSH terms 'human', 'blood cells' and 'transcription factors'. From this search 2,000 abstracts were selected and hand annotated according to a small taxonomy of 48 classes based on a chemical classification. Among the classes, 36 terminal classes were used to annotate the GENIA corpus.

The GENIA corpus is important for two major reasons: the first is that it provides the largest single source of annotated training data for the NE task in molecular biology and the second is in the breadth of classification. Although 36 classes is a fraction of the classes contained in major taxonomies it is still the largest class set that has been attempted so far for the NE task. In this respect it is an important test of the limits of human and machine annotation capability. For the shared task we decided however to simplify the 36 classes and used only the classes *protein*, *DNA*, *RNA*, *cell line* and *cell type*. The first three incorporate several subclasses from the original taxonomy while the last two are interesting in order to make the task realistic for post-processing by a potential template filling application.

For testing purposes we used a newly annotated collection of MEDLINE abstracts from the GENIA project. 404 abstracts were used that were annotated for the same classes of entities: Half of them were from the same domain as the training data and the other half of them were from the super-domain of 'blood cells' and 'transcription factors'. Our hope was that this should provide an important test of generalizability in the methods used.

3 Evaluation

The 2,000 abstracts of the GENIA corpus version 3.02 which had already been made publicly available were formatted for IOB2 notation and made available as training materials. For testing, additional 404 abstracts were randomly selected from an unpublished set of the GENIA corpus and the annotations were re-checked by a biologist. The training set consists of abstracts retrieved from the MEDLINE database with MeSH terms 'human', 'blood cells' and 'transcription factors', and their publication year ranges over 1990∼1999. Most parts of the test set include abstracts retrieved with the same set of MeSH terms, and their publication year ranges over 1978∼2001. To see the effect of publication year, the test set was roughly divided into four subsets: **1978-1989 set** (which represents an old age from the viewpoint of the models that will be trained using the training set), **1990-1999 set** (which represents the same age as the training set), **2000-2001 set** (which represents a new age compared to the training set) and **S/1998-2001 set** (which represents roughly a new age in a super domain). The last subset represents a super domain and the abstracts was retrieved with MeSH terms, 'blood cells' and 'transcription factors' (without 'human')[2]. Table 1 illustrates the size of the data sets

Table 2 shows the number of entities annotated in each data set[3]. As seen in the table, the annotation density of proteins increases over the ages significantly, whereas the annotation density of DNAs and RNAs increases in the **1990-1999 set** and slightly decreases in the **2000-2001 set**. This tendency roughly corresponds to the expansion in the subject area as a whole that can be estimated from statistics on the MeSH terms introduced in each age shown in Table 3. This observation suggests that the density of mention of a class of entities in academic papers is affected by the amount of interest the entity receives in each age.

Figure 2 shows the ratio of annotated structures in each set. In accordance with our expectation, the **1990-1999 set** has the most similar annotation trait with the training set. The **2000-2001 set** is also similar to the training set, but the **1978-1989 set** had quite a different distribution of entity classes. The variation of domain does not seem to make any significant difference to the distribution of entities mentioned. One reason may be the large fraction of abstracts from the same domain in the super domain set. In fact, among 206 abstracts in the super domain set, 140 abstracts (69%) are also from the same domain. It also corresponds to the fraction in the whole MEDLINE database: among 9,362 abstracts that can be retrieved with MeSH terms, 'blood cells' and 'transcription factors', 6,297 abstracts (67%) can also be retrieved with MeSH terms 'human', 'blood cells' and 'transcription factors'.

To simplify the annotation task to a simple linear sequential analysis problem, embedded structures have been removed leaving only the

[2]The **S/1998-2001 set** includes the whole **2000-2001 set**.

[3]The figures in the parenthesis are the average number of entities per an abstract in each set.

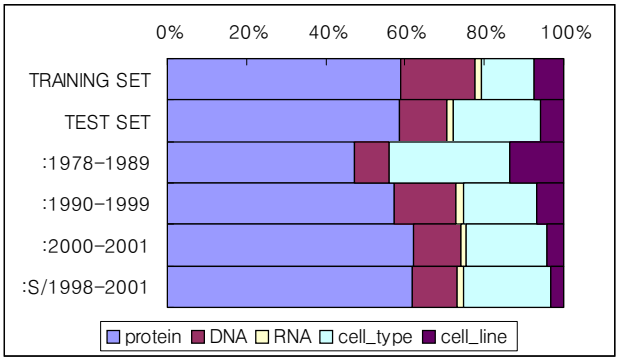
	# abs	# sentences	# words
Training Set	2,000	20,546 (10.27/abs)	472,006 (236.00/abs) (22.97/sen)
Test Set	404	4,260 (10.54/abs)	96,780 (239.55/abs) (22.72/sen)
1978-1989	104	991 (9.53/abs)	22,320 (214.62/abs) (22.52/sen)
1990-1999	106	1,115 (10.52/abs)	25,080 (236.60/abs) (22.49/sen)
2000-2001	130	1,452 (11.17/abs)	33,380 (256.77/abs) (22.99/sen)
S/1998-2001	206	2,270 (11.02/abs)	51,957 (252.22/abs) (22.89/sen)

Table 1: Basic statistics for the data sets

	protein	DNA	RNA	cell_type	cell_line	ALL
Training Set	30,269 (15.1)	9,533 (4.8)	951 (0.5)	6,718 (3.4)	3,830 (1.9)	51,301 (25.7)
Test Set	5,067 (12.5)	1,056 (2.6)	118 (0.3)	1,921 (4.8)	500 (1.2)	8,662 (21.4)
1978-1989	609 (5.9)	112 (1.1)	1 (0.0)	392 (3.8)	176 (1.7)	1,290 (12.4)
1990-1999	1,420 (13.4)	385 (3.6)	49 (0.5)	459 (4.3)	168 (1.6)	2,481 (23.4)
2000-2001	2,180 (16.8)	411 (3.2)	52 (0.4)	714 (5.5)	144 (1.1)	3,501 (26.9)
S/1998-2001	3,186 (15.5)	588 (2.9)	70 (0.3)	1,138 (5.5)	170 (0.8)	5,152 (25.0)

Table 2: Absolute (and relative) frequencies for NEs in each data set. Figures for the test set are broken down according to the age of the data.

Figure 2: Ratio of annotated NEs



outermost structures (i.e. the longest tag sequence). Consequently, a group of coordinated entities involving ellipsis are annotated as one structure like in the following example:

> ... in [lymphocytes] and [T- and B-lymphocyte] count in ...

In the example, *"T- and B-lymphocyte"* is annotated as one structure but involves two entity names, *"T-lymphocyte"* and *"B-lymphocyte"*, whereas *"lymphocytes"* is annotated as one and involves as many entity names.

	prot.	DNA	RNA	ctype	cline
1978-1989	3.3	2.8	1.5	0.1	0
1990-1999	16.6	4.6	7.4	0.2	1.5
2000-2001	40.0	4.0	3.5	0	0

Table 3: MeSH terms in each age (#/year)

4 Evaluation Methodology

Results are given as F-scores using a modified version of the CoNLL evaluation script and are defined as $F = (2PR)/(P + R)$, where P denotes Precision and R Recall. P is the ratio of the number of correctly found NE chunks to the number of found NE chunks, and R is the ratio of the number of correctly found NE chunks to the number of true NE chunks. The script outputs three sets of F-scores according to exact boundary match, right and left boundary matching. In the right boundary matching only right boundaries of entities are considered without matching left boundaries and vice versa.

5 Participating Systems

5.1 Classification Models

Roughly four types of classification models were applied by the eight participating systems; Support Vector Machines (SVMs), Hidden Markov Models (HMMs), Maximum Entropy Markov Models (MEMMs) and Conditional Random Fields (CRFs). The most frequently applied

models were SVMs with totally five systems adopting SVMs as the classification models either in isolation (Park et al., 2004; Lee et al., 2004) or in combination with other models (Zhou and Su, 2004; Song et al., 2004; Rössler, 2004). HMMs were employed by one system in isolation (Zhao, 2004) and by two systems in combination with SVMs (Zhou and Su, 2004; Rössler, 2004). Similarly, CRFs were employed by one system in isolation (Settles, 2004) and by another system in combination with SVMs (Song et al., 2004). It is somewhat surprising that Maximum Entropy Models were applied by only one system (Finkel et al., 2004), while it was the most successfully applied model in the CoNLL-2003 Shared Task of Named Entity Recognition, and at this time also the MEMM system yields quite good performance. One interpretation on this may be the CRF is often regarded as a kind of version-upped model of the MEMM (in the sense that both are conditional, exponential models) and thus is replacing MEMM.

5.2 Features and External Resources

It has been found that utilizing various sources of information is crucial to get good performance in this kind of task. Table 4 outlines some of the features exploited by the systems participating in the JNLPBA 2004 shared task (the table also lists the classification models employed and external resources exploited by the systems to provide the outline of the systems at a glance).

Lexical features (words) were widely exploited by three systems that didn't employ SVMs. It seems that this may be due to SVMs' high time complexity and actually other two SVM systems also employed lexical features only in a limited way. Instead, affixes, orthographic features or word shapes that are all generalized forms of lexical features were actively exploited by most of the systems. The ATCG sequence feature is an example of domain specific orthographic features and was incorporated in three systems. Park et al. (2004) suggested the use of word variation features, a unique way of selecting substrings from words, but the effectiveness was not reported.

Part-of-speech information was incorporated in five systems: four of them utilized domain-specialized part-of-speech taggers (Zhou and Su, 2004; Finkel et al., 2004; Song et al., 2004; Park et al., 2004) and the other utilized general-purpose taggers (Lee et al., 2004). BeseNP tags and deep syntactic features were also exploited by several systems but the effectiveness was not clearly examined.

The top-ranked two systems incorporated information from gazetteers and employed abbreviation handling mechanisms, which were reported to give good effect. However, one participant (Settles, 2004) reported that their attempt to utilize gazetteers (together with other resources) had failed in gaining better overall performance.

To overcome the shortage of training materials, several systems attempted to use external resources. Gazetteers are also examples of such resources. MEDLINE database was explored as a source of a large corpus that is similar to the training corpus, but one participant (Rössler, 2004) reported the attempt was not successful. Finkel et al. (2004) exploited BNC corpus and World Wide Web as knowledge sources and achieved good performance., but the effectiveness of the use of such resources was not clearly examined. Song et al. (2004) exploited automatically generated virtual examples and reported good effect on both recall and precision. Lee et al. (2004) utilized external protein and gene taggers instead of using gazetteers but the effectiveness was not reported.

5.3 Performances[4]

Table 5 lists entity recognition performance of each system on each test set. The baseline model (BL) utilizes lists of entities of each class collected from the training set, and performs longest match search for entities through the test set. Frequency of each entity with each class is referred to break ties.

It may be notable that SVMs worked much better in combination with other models, while other models showed reasonable performance even in isolation. This fact suggests that global optimization over whole sequence (e.g, Viterbi optimization) is crucial in this type of tasks. As is well known, the outputs of SVMs are not easy to use in global optimization. It seems (Zhou and Su, 2004) overcomes the drawback of SVMs by mapping the SVM output into probability, and complementing it with Markov models. Their remarkable performance seems due to the well designed classification model and the

[4] A comprehensive report of systems performance is available at http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html.

	CM	lx	af	or	sh	gn	wv	ln	gz	po	np	sy	tr	ab	ca	do	pa	pr	ext.
Zho	SH	-	+	+	-	+	-	-	+	+	-	-	+	+	+	-	-	+	-
Fin	M	+	+	-	+	-	-	-	+	+	-	+	-	+	-	+	+	+	B, W
Set	C	+	+	+	+	-	-	-	(+)	-	-	-	(+)	-	-	-	-	+	(W)
Son	SC	*	+	+	-	-	-	-	-	+	+	-	-	-	-	-	-	+	V
Zha	H	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	M
Rös	SH	-	+	+	-	+	-	+	-	-	-	-	-	-	-	-	-	+	(M)
Par	S	-	+	+	+	+	+	-	-	+	+	-	+	-	-	-	-	-	M, P
Lee	S	*	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	Y, G

Table 4: Overview of participating systems in terms of classification models, main features and external resources, sorted by performance. Classification Model (CM): S: SVM; H: HMM; M: MEMM; C: CRF; lx: lexical features; af: affix information (character n-grams); or: orthographic information; sh: word shapes; gn: gene sequences (ATCG sequences); wv: word variations; ln: word length; gz: gazetteers; po: part-of-speech tags; np: noun phrase tags; sy: syntactic tags; tr: word triggers; ab: abbreviations; ca: cascaded entities; do: global document information; pa: parentheses handling; pr: previously predicted entity tags; External resources (ext): B: British National Corpus; M: MEDLINE corpus; P: Penn Treebank II corpus; W: world wide web; V: virtually generated corpus; Y: Yapex; G: GAPSCORE.

	1978-1989 set	1990-1999 set	2000-2001 set	S/1998-2001 set	Total
Zho	75.3 / 69.5 / 72.3	77.1 / 69.2 / 72.9	75.6 / 71.3 / 73.8	75.8 / 69.5 / 72.5	76.0 / 69.4 / 72.6
Fin	66.9 / 70.4 / 68.6	73.8 / 69.4 / 71.5	72.6 / 69.3 / 70.9	71.8 / 67.5 / 69.6	71.6 / 68.6 / 70.1
Set	63.6 / 71.4 / 67.3	72.2 / 68.7 / 70.4	71.3 / 69.6 / 70.5	71.3 / 68.8 / 70.1	70.3 / 69.3 / 69.8
Son	60.3 / 66.2 / 63.1	71.2 / 65.6 / 68.2	69.5 / 65.8 / 67.6	68.3 / 64.0 / 66.1	67.8 / 64.8 / 66.3
Zha	63.2 / 60.4 / 61.8	72.5 / 62.6 / 67.2	69.1 / 60.2 / 64.7	69.2 / 60.3 / 64.4	69.1 / 61.0 / 64.8
Rös	59.2 / 60.3 / 59.8	70.3 / 61.8 / 65.8	68.4 / 61.5 / 64.8	68.3 / 60.4 / 64.1	67.4 / 61.0 / 64.0
Par	62.8 / 55.9 / 59.2	70.3 / 61.4 / 65.6	65.1 / 60.4 / 62.7	65.9 / 59.7 / 62.7	66.5 / 59.8 / 63.0
Lee	42.5 / 42.0 / 42.2	52.5 / 49.1 / 50.8	53.8 / 50.9 / 52.3	52.3 / 48.1 / 50.1	50.8 / 47.6 / 49.1
BL	47.1 / 33.9 / 39.4	56.8 / 45.5 / 50.5	51.7 / 46.3 / 48.8	52.6 / 46.0 / 49.1	52.6 / 43.6 / 47.7

Table 5: Performance of each participating system and a baseline model (BL) (recall / precision / F-score)

rich set of features.

As is naturally expected, most systems (5 out of 8) show their best performance on the **1990-1999 set** which is believed to have the most similar annotation trait. The same tendency is observed more clearly with recall (7 out of 8 show their best performance on the **1990-1999 set**) while no such tendency is observed with precision. If we accept such tendency of showing best performance on the most similar test set as natural, one interpretation on the observation might be that positive information has been well exploited while negative information has not. Clearly, a such case is the baseline model which utilizes only positive information and no negative information. Finkel et al. (2004) explicitly pointed out the problem of "abusing" positive information with regard to using gazetteers, and utilized frequency information from BNC corpus to prevent such "abusement". Settles (2004)'s CRF system deserves special note in the sense that it achieved comparable performance to top ranked systems with a rather simple feature set. This fact may suggest that integration of information is as much important as development of useful features.

As the resulting performance may not seem very successful, other systems suggest interesting approaches: Song et al. (2004) reports about the effectiveness of using virtual examples. Zhao (2004) reports about the usefulness of unlabeled MEDLINE corpus as a complement to expensive and limited size of labeled corpus. Rössler (2004) reports their experience to adapt an NER system for German to biomedical domain. Park et al. (2004) reports their efforts to find out useful information by corpus comparison. Lee et al. (2004) suggests the use of external protein/gene taggers instead of using gazetteers.

6 Conclusion

While it is not entirely meaningful to rank systems performance according to simple F-scores,

the accuracy results do nevertheless show some important trends that may help guide future system developers in the bio-entity task. It is clear that we have to move beyond simple lexical features if we want to obtain high levels of performance in molecular biology and the top performing systems were seen to be those that employed strong learning models (SVM, MEMM and CRF), rich feature sets, support for 'difficult' constructions such as parenthesized expressions and a sophisticated mix of external resources such as gazette lists and ontologies which provide terminological resources. It is also interesting to observe that we have seen the beginning of a trend in the use of the Web which can provide online access to dynamically updated resources or sophisticated search for sets of similar terms.

7 Acknowledgements

We gratefully acknowledge Prof. Jun'ichi Tsujii, University of Tokyo, for his generous support of the shared task. The GENIA corpus is a product of the GENIA project which is supported by the Information Mobility Project (CREST, JST) and the Genome Information Science Project (MEXT).

References

Jin-Dong Kim, Tomoko Ohta, Yuka Tateishi, and Jun'ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl.1):180–182.

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

Erik F. Tjong Kim Sang and Fien 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.

C. J. van Rijsbergen. 1979. *Information Retrieval*. Butterworths, London.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi and Jun'ichi 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.

GuoDong Zhou and Jian 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.

Jenny Finkel, Shipra Dingare, Huy Nguyen, Malvina Nissim, Gail Sinclair and Christopher 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.

Burr 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.

Yu Song, Eunju Kim, Gary Geunbae Lee and Byoung-kee 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.

Shaojun 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.

Marc Rössler. 2004. Adapting an 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.

Kyung-Mi Park, Seon-Ho Kim, Do-Gil Lee and Hae-Chang 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.

Chih Lee, Wen-Juan Hou and Hsin-Hsi Chen. 2004. Annotating Multiple Types of Biomedical Entities: A Single Word Classificication Approach. in *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004)*, Geneva, Switzerland.