Comparison between Tagged Corpora for the Named Entity Task

Chikashi NOBATA
Kansai Advanced Research Center
Communications Research Laboratory
588-2 Iwaoka, Iwaoka-cho, Nishi-ku
Kobe, Hyogo, 651-2492 JAPAN
nova@crl.go.jp

Nigel COLLIER and Jun'ichi TSUJII
Department of Information Science
Graduate School of Science
University of Tokyo, Hongo 7-3-1
Bunkyo-ku, Tokyo, 113-0033 JAPAN
{nigel, tsujii}@is.s.u-tokyo.ac.jp

Abstract

We present two measures for comparing corpora based on information theory statistics such as gain ratio as well as simple term-class frequency counts. We tested the predictions made by these measures about corpus difficulty in two domains — news and molecular biology — using the result of two well-used paradigms for NE, decision trees and HMMs and found that gain ratio was the more reliable predictor.

1 Introduction

With the advent of the information society and increasing availability of large amounts of information in electronic form, new technologies such as information extraction are emerging to meet user's information access needs. Recent evaluation conferences such as TREC (Voorhees and Harman, 2000) showed the feasibility of this task and highlighted the need to combine information retrieval (IR) and extraction (IE) to go beyond simply offering the user a long ranked list of interesting documents to providing facts for user's questions.

The problem of domain dependence remains a serious one and in fact there has been very little work so far to compare the difficulty of IE tasks for different domains and their corpora. Such knowledge is useful for developing IE systems that are portable between domains. This paper begins to address this issue, in particular the lowest level of IE task, defined in the TIPSTER sponsored MUC-6 conference (MUC, 1995) as named entity (NE). This is emerging as a key technology in several other IE-related tasks such as question answering. We seek here to show theoretically motivated measures for comparing the difficulty of corpora for the NE task in two domains, newswire and molecular-biology. We then test the predictions made by these measures against actual system performance.

Recently IE systems based on supervised learning paradigms such as hidden Markov models (Bikel et al., 1997), maximum entropy (Borthwick et al., 1998) and decision trees (Sekine et al., 1998) have emerged that should be easier to adapt to new domains than the dictionary-based systems of the past. Much of this work has taken advantage of smoothing techniques to overcome problems associated with data sparseness (Chen and Goodman, 1996).

The two corpora we use in our NE experiments represent the following domains:

- Newswire: acquisition of names of people, organizations and monetary units etc., from the MUC-6 data set.
- Molecular-biology: acquisition of proteins, DNAs, RNAs etc. from a subset of the MEDLINE database (MEDLINE, 1999).

Information extraction in the molecular-biology domain (Sekimizu et al., 1998) (Craven and Kumlien, 1999) (Rindflesch et al., 2000) has recently become a topic of interest to the NLP community. This is a result of the need to formalise the huge number of research results that appear in free-text form in online collections of journal abstracts and papers such as MEDLINE for databases such as Swissprot (Bairoch and Apweiler, 1997) and also to search such collections for facts in an intelligent way.

The purpose of our study is not to show a high level of absolute system performance. In fact since we use only the MUC-6 executive succession data set of 60 articles and a new MEDLINE data set of 100 articles we cannot hope to achieve performance limits. What we aim to do is to compare model performance against the predictions of corpus difficulty made by two different methods. In the rest of this paper we firstly introduce the NE models used for evaluation, the two corpora we
examined and then the difficulty comparison metrics. Predictive scores from the metrics are examined against the actual performance of the NE models.

2 Models

Recent studies into the use of supervised learning-based models for the NE task in the molecular-biology domain have shown that models based on hidden Markov models (HMMs) (Collier et al., 2000) and decision trees (Nobata et al., 1999) are not only adaptable to this highly technical domain, but are also much more generalizable to new classes of words than systems based on traditional hand-built heuristic rules such as (Fukuda et al., 1998). We now describe two models used in our experiments based on the decision trees package C4.5 (Quinlan, 1993) and HMMs (Rabiner and Juang, 1986).

2.1 Decision tree named entity recogniser: NE-DT

A decision tree is a type of classifier which has "leaf nodes" indicating classes and "decision nodes" that specify some test to be carried out, with one branch or subtree for each possible outcome of the test. A decision tree can be used to classify an object by starting at the root of the tree and moving through it until a leaf is encountered. When we can define suitable features for the decision tree, the system can achieve good performance with only a small amount of training data.

The system we used is based on one that was originally created for Japanese documents (Sekine et al., 1998). It has two phases, one for creating the decision tree from training data and the other for generating the class-tagged text based on the decision tree. When generating decision trees, trigrams of words were used. For this system, words are considered to be quadruple features. The following features are used to generate conditions in the decision tree:

Part-of-speech information: There are 45 part-of-speech categories, whose definitions are based on Pennsylvania Treebank's categories. We use a tagger based on Adwait Ratnaparkhi's method (Ratnaparkhi, 1996).

Character type information: Orthographic information is considered such as upper case, lower case, capitalization, numerical expressions, symbols. These character features are the same as those used by NEHMM described in the next section and shown in Table 1.

Word lists specific to the domain: Word lists are made from the training corpus. Only the 200 highest frequency words are used.

2.2 Hidden Markov model named entity recogniser: NEHMM

HMMs are a widely used class of learning algorithms and can be considered to be stochastic finite state machines. In the following model, summarized here from the full description given in (Collier et al., 2000), we consider words to be ordered pairs consisting of a surface word, $W$, and a word feature, $F$, given as $< W, F >$. The word features themselves are discussed below. As is common practice, we need to calculate the probabilities for a word sequence for the first word's name class and every other word differently since we have no initial name-class to make a transition from. Accordingly we use the following equation to calculate the initial name class probability,

$$\Pr(NC_1 | < W_1, F_1 >) = \sigma_0 f(NC_{first} | < W_1, F_1 >) + \sigma_1 f(NC_{first} | < ., F_1 >) + \sigma_2 f(NC_{first})$$

(1)

and for all other words and their name classes as follows:

$$\Pr(NC_t | < W_t, F_t >, < W_{t-1}, F_{t-1 }>, NC_{t-1}) = \lambda_0 f(NC_t | < W_t, F_t >, < W_{t-1}, F_{t-1 }>, NC_{t-1}) + \lambda_1 f(NC_t | < ., F_t >, < W_{t-1}, F_{t-1 }>, NC_{t-1}) + \lambda_2 f(NC_t | < W_t, F_t >, < ., F_{t-1 }>, NC_{t-1}) + \lambda_3 f(NC_t | < ., F_t >, < ., F_{t-1 }>, NC_{t-1}) + \lambda_4 f(NC_t | NC_{t-1}) + \lambda_5 f(NC_t)$$

(2)

where $f()$ is calculated with maximum-likelihood estimates from counts on training data.

In our current system we set the constants $\lambda_i$ and $\sigma_i$ by hand and let $\sum \sigma_i = 1.0$, $\sum \lambda_i = 1.0$, $\sigma_0 / \sigma_1 \geq \sigma_2$, $\lambda_0 / \lambda_1 \ldots \geq \lambda_5$. The current name-class $NC_t$ is conditioned on the current word and feature, the previous name-class, $NC_{t-1}$, and previous word and feature.

Equations 1 and 2 implement a linear-interpolating HMM that incorporates a number of sub-models designed to reduce the effects of data sparseness.
Table 1: Word features with examples

| Word Feature          | Example | Feature | Ex.  |
|-----------------------|---------|---------|------|
| TwoDigitNumber        | 25      | OpenSquare |      |
| FourDigitNumber       | 2000    | CloseSquare |    |
| DigitNumber           | 15012   | Colon   |      |
| SingleCap             | M       | SemiColon |     |
| GreekLetter           | alpha   | Percent  | %    |
| CapsAndDigits         | T2      | OpenParen | (    |
| TwoCaps               | RalGDS  | CloseParen | )   |
| LettersAndDigits      | p52     | Comma   |      |
| InitCap               | Interleukin | FullStop |     |
| LowCaps               | kappaB  | Determiner | the |
| Lowercase             | kinases | conjunction and | |
| Hyphon                | -       | Other *+# | |
| Backslash             | /       |         |      |

Once the state transition probabilities have been calculated according to Equations 1 and 2, the Viterbi algorithm (Viterbi, 1967) is used to search the state space of possible name class assignments in linear time to find the highest probability path, i.e. to maximise $Pr(W, NC)$. The final stage of our algorithm that is used after name-class tagging is complete is to use a clean-up module called Unity. This creates a frequency list of words and name-classes and then re-tags the text using the most frequently used name class assigned by the HMM. We have generally found that this improves F-score performance by between 2 and 4%, both for re-tagging spuriously tagged words and for finding untagged words in unknown contexts that had been correctly tagged elsewhere in the text.

Table 1 shows the character features that we used in both NEHMM and NE-DT. Our intuition is that such features will help the model to find similarities between known words that were found in the training set and unknown words and so overcome the unknown word problem.

3 Corpora

We used two corpora in our experiments representing two popular domains in IE, molecular-biology (from MEDLINE) and newswire texts (from MUC-6). These are now described.

3.1 MUC-6

The corpus for MUC-6 (MUC, 1995) contains 60 articles, from the test corpus for the dry and formal runs. An example can be seen in Figure 1. We can see several interesting features of the domain such as the focus of NEs on people and organization profiles. Moreover we see that there are many pre-name clue words such as "Ms." or "Rep." indicating that a Republican politician’s name should follow.

3.2 Biology

In our tests in the domain of molecular-biology we are using abstracts available from PubMed’s MEDLINE. The MEDLINE database is an online collection of abstracts for published journal articles in biology and medicine and contains more than nine million articles. Currently we have extracted a subset of MEDLINE based on a search using the keywords human AND blood cell AND transcription factor yielding about 3650 abstracts. Of these 100 documents were NE tagged for our experiments using a human domain expert. An example of the annotated abstracts is shown in Figure 2. In contrast to MUC-6 each article is quite short and there are few pre-class clue words making the task much more like terminology identification and classification than pure name finding.

4 A first attempt at corpus comparison based on simple token frequency

A simple and intuitive approach to NE task difficulty comparison used in some previous studies such as (Palmer and Day, 1997) who studied corpora in six different languages, compares class to term-token ratios on the assumption that rarer classes are more difficult to acquire. The relative frequency counts from these ratios also give an indirect measure of the granularity of a class, i.e. how wide it is. While this is appealing, we show that this approach does not necessarily give the best metric for comparison.

Tables 2 and 3 show the ratio of the number of different words used in NEs to the total number of words in the NE class vocabulary. The number of different tokens is influenced by the corpus size and is not a suitable index that can uniformly show the difficulty for different NE tasks, therefore it should be normalized. Here we use words as tokens. A value close to zero indicates little variation within the class and should imply that the class is easier to acquire. We see that the NEs in the biology domain seem overall to be easier to acquire than those in the MUC-6 domain given lexical variation.

The figures in the second columns of Tables 2 and 3 are normalized so that all numerals are replaced by a single token. It still seems though that MUC-6 is a considerably more challenging domain than biology. This is despite the fact that the ratios for ENAMEX expressions such as Date,
A graduate of Harvard Law School, Ms. Washington worked as a lawyer for the corporate finance division of the SEC in the late 1970s. She has been a congressional staffer since 1979. Separately, transition officials said that Frank Newman, vice chairman and chief financial officer of BankAmerica Corp., is expected to be nominated as assistant Treasury secretary for domestic finance.

Figure 1: Example sentences taken from the annotated MUC-6 NE text

Table 2: Frequency values for words in the MUC-6 test corpus

| Class    | Original         | Norm. numerals |
|----------|------------------|----------------|
| Org.     | 0.28 (507/1783)  | 0.28 (507/1783) |
| Person   | 0.45 (391/858)   | 0.42 (381/838)  |
| Loc.     | 0.38 (148/390)   | 0.38 (148/390)  |
| Date     | 0.33 (123/452)   | 0.11 (60/542)   |
| Time     | 1.00 (3/3)       | 1.00 (3/3)      |
| Money    | 0.33 (138/423)   | 0.05 (20/422)   |
| Percent  | 0.39 (42/108)    | 0.03 (3/108)    |
| All      | 0.33 (1342/4087) | 0.27 (1122/4087) |

Money and Percent all fall significantly. Expressions in the Time class are so rare however that it is difficult to make any sort of meaningful comparison. In the biology corpus, the ratios are not significantly changed and the NE classes defined for biology documents seem to have the same characteristics as non-numeric ENAMEX classes in MUC-6 documents.

Comparing between the biology documents and the MUC-6 documents, we may say that identifying entities in biology documents is easier than identifying ENAMEX entities in MUC-6 documents.

5 Experiments

We evaluated the performance of our two systems using a cross validation method. For the MUC-6 corpus, 6-fold cross validation was performed on the 60 texts and 5-fold cross validation was performed on the 100 texts in the biology corpus.

We use "F-scores" for evaluation of our experiments (Van Rijsbergen, 1979). "F-score" is a measurement combining "Recall" and "Precision" and defined in Equation 3. "Recall" is the percentage of correct answers among the answers proposed by the system. The F-scores presented here are automatically calculated using a scoring program (Chinchor, 1995).

\[
F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}
\]

In Table 4 we show the actual performance of our term recognition systems, NE-DT and NEHMM. We can see that corpus comparisons based only on class-token ratios are inadequate to explain why both systems' performance was about the same in both domains or why NEHMM did better in both test corpora than NE-DT. The difference in performance is despite there being more training examples in biology (3301 NEs) than in MUC-6 (2182 NEs). Part of the reason for this is
Table 4: Performance of the NE systems

| System                | MUC-6  | Biology |
|----------------------|--------|---------|
| NEHMM with Unity     | 78.4   | 75.0    |
| NEHMM w/o Unity      | 74.2   | 73.1    |
| NE-DT                | 68.5   | 69.4    |

That the class-token ratios ignore individual system knowledge, i.e. the types of features that can be captured and useful in the corpus domain. Among other considerations they also fail to consider the overlap of words and features between classes in the same corpus domain.

6 Corpus comparison based on information theoretical measures

In this section we attempt to present measures that overcome some of the limitations of the class-token method. We evaluate the contribution from each feature used in our NE recognition systems by calculating its entropy. There are three types of feature information used by our two systems: lexical information, character type information, and part-of-speech information.

The entropy for NE classes \(H(C)\) is defined by

\[
H(C) = - \sum_{c \in C} p(c) \log_2 p(c)
\]

where:

\[
p(c) = \frac{n(c)}{N}
\]

\(n(c)\): the number of words in class \(c\)

\(N\): the total number of words in text

We can calculate the entropy for features in the same way.

When a feature \(F\) is given, the conditional entropy for NE classes \(H(C|F)\) is defined by

\[
H(C|F) = - \sum_{c \in C} \sum_{f \in F} p(c, f) \log_2 p(c|f)
\]

where:

\[
p(c, f) = \frac{n(c, f)}{N}
\]

\[
p(c|f) = \frac{n(c, f)}{n(f)}
\]

\(n(c)\): the number of words in class \(c\) with the feature value \(f\)

\(n(f)\): the number of words with the feature value \(f\)

Using these entropies, we can calculate information gain (Breiman et al., 1984) and gain ratio (Quinlan, 1990). Information gain for NE classes and a feature \(I(C; F)\) is given as follows:

\[
I(C; F) = H(C) - H(C|F)
\]

The information gain \(I(C; F)\) shows how the feature \(F\) is related with NE classes \(C\). When \(F\) is completely independent of \(C\), the value of \(I(C; F)\) becomes the minimum value 0. The maximum value of \(I(C; F)\) is equivalent to that of \(H(C)\), when the feature \(F\) gives sufficient information to recognize named entities. Information gain can also be calculated by:

\[
I(C; F) = H(C) + H(F) - H(C, F)
\]

Gain ratio is the normalized value of information gain. The gain ratio \(GR(C; F)\) is defined by

\[
GR(C; F) = \frac{I(C; F)}{H(C)}
\]

The range of the gain ratio \(GR(C; F)\) is \(0 \leq GR(C; F) \leq 1\) even when the class entropy is different in various corpora, so we can compare the values directly in the different NE recognition tasks.

6.1 Character types

Character type features are used to identify named entities in the MUC-6 and biology corpus. However, the distribution of the character types are quite different between these two types of documents as we can see in Table 5. We see through the gain-ratio score that character type information has a greater predictive power for classes in MUC-6 than biology due to the higher entropy of character type and class sequences in the biology corpus, i.e. the greater disorder of this information. The result partially shows why identification and classification is harder in biological documents than in newspaper articles such as the MUC-6 corpus.

6.2 Part-of-speech

Table 6 shows the entropy scores for part-of-speech (POS) sequences in the two corpora. We see through the gain ratio scores that POS information is not so powerful for acquiring NEs in the biology domain compared to the MUC-6 domain.
Table 5: Values of Entropy for character type

| Entropy | MUC-6     | Biology   |
|---------|-----------|-----------|
| H(T)    | 1.880     | 2.013     |
| H(C)    | 0.890     | 1.264     |
| H(C,T)  | 2.345     | 2.974     |
| GR(C,T) | 0.478     | 0.239     |

Table 6: Values of Entropy for POSs

| Entropy | MUC-6     | Biology   |
|---------|-----------|-----------|
| H(P)    | 4.287     | 4.037     |
| H(C)    | 0.890     | 1.264     |
| H(C,P)  | 4.750     | 5.029     |
| I(C;P)  | 0.426     | 0.272     |
| GR(C;P) | 0.479     | 0.216     |

In fact POS information for biology is far less useful than character information when we compare the results in Tables 5 and 6, whereas POS has about the same predictive power as character information in the MUC-6 domain. One likely explanation for this is that the POS tagger we use in NE-DT is trained on a corpus based on newspaper articles, therefore the assigned POS tags are often incorrect in biology documents.

6.3 Lexical information

Table 7 shows the entropy statistics for the two domains. Although entropy for words in biology is lower than MUC-6, the entropy for classes is higher leading to a lower gain ratio in biology. We also note that, as we would expect, in comparison to the other two types of knowledge, surface word forms are by far the most useful type of knowledge with a gain ratio in MUC-6 of 0.897 compared to 0.479 for POS and 0.478 for character types in the same domain. However, such knowledge is also the least generalizable and runs the risk of data-sparseness. It therefore has to be complemented by more generalizable knowledge such as character features and POS.

6.4 Comparison between the combination of features

In this section we show a comparison of gain ratio for the features used by both systems in each corpus. Values of gain ratio for each feature set are shown on the ‘GR’ column in Tables 8, 9, 10, and 11. The values of GR show that surface words have the best contribution in both corpora for both systems. We can see that gain ratio for all features in NE-DT is actually lower than the top level model for NEHMM in biology, reflecting the actual system performance that we observed.

Values for cross entropy between training and test sets are shown in Tables 8, 9, 10, and 11 to show the combination of gain ratio for the features used by both systems in each corpus. Values of cross entropy for each feature set are shown on the ‘Features’ column. ‘ALL’ in Tables 10 and 11 means all the features used in decision tree, i.e. \( \{F_{t-1}, d_{t+1}, F_{t-1,d_{t+1}} \} \).
Table 10: Values of Entropy for NE-DT features in the MUC-6 corpus

| GR       | Cross Entropy | Coverage | Features   |
|----------|---------------|----------|------------|
| 0.989    | 1.29(1.38-1.77) | 0.12(0.10-0.13) | ALL        |
| 0.402    | 5.25(0.99-5.32) | 1.00(0.99-1.00) | F           |
| 0.458    | 5.57(0.51-2.87) | 1.00(0.99-1.00) | F           |
| 0.944    | 7.56(7.97-7.57) | 0.81(0.65-0.83) | W           |
| 0.678    | 5.17(7.57-1.97) | 0.68(0.96-0.98) | F<sub>1</sub>,<sub>2</sub> |
| 0.969    | 3.97(7.67-4.07) | 0.99(0.99-1.00) | F<sub>1</sub>,<sub>2</sub> |
| 0.977    | 5.74(4.10-4.88) | 0.36(0.34-0.40) | W<sub>1</sub>,<sub>2</sub> |
| 0.822    | 3.95(7.90-9.40) | 0.69(0.87-0.91) | F<sub>1</sub>,<sub>1</sub>,<sub>1</sub>,<sub>1</sub> |
| 0.807    | 4.99(7.47-5.08) | 0.96(0.95-0.95) | F<sub>1</sub>,<sub>1</sub>,<sub>1</sub>,<sub>1</sub> |
| 0.989    | 1.89(1.67-2.16) | 0.15(0.13-0.17) | W<sub>1</sub>,<sub>1</sub>,<sub>1</sub>,<sub>1</sub> |

Table 11: Values of Entropy for NE-DT features in the biology corpus

| GR       | Cross Entropy | Coverage | Features   |
|----------|---------------|----------|------------|
| 0.937    | 3.31(5.00-2.50) | 0.18(0.15-0.19) | ALL        |
| 0.237    | 5.31(5.21-5.38) | 1.00(0.99-1.00) | F           |
| 0.297    | 3.27(3.14-3.41) | 1.00(1.00-1.00) | F           |
| 0.418    | 7.64(7.65-7.79) | 0.87(0.82-0.88) | W           |
| 0.370    | 7.78(7.69-7.85) | 0.97(0.96-0.97) | F<sub>1</sub>,<sub>2</sub> |
| 0.383    | 4.57(4.28-4.67) | 0.98(0.98-0.99) | F<sub>1</sub>,<sub>2</sub> |
| 0.586    | 5.71(5.37-5.93) | 0.48(0.45-0.50) | W<sub>1</sub>,<sub>2</sub> |
| 0.541    | 8.02(8.25-9.02) | 0.88(0.87-0.90) | F<sub>1</sub>,<sub>1</sub>,<sub>1</sub>,<sub>1</sub> |
| 0.592    | 4.65(5.36-5.54) | 0.92(0.94-0.96) | F<sub>1</sub>,<sub>1</sub>,<sub>1</sub>,<sub>1</sub> |
| 0.764    | 2.56(2.25-2.76) | 0.20(0.17-0.21) | W<sub>1</sub>,<sub>1</sub>,<sub>1</sub>,<sub>1</sub> |

In this paper we have shown that measures for comparing corpora based just on class-token ratios have difficulty predicting system performance and cannot adequately explain the difficulty of the NE task either generally or for specific systems.

We should be cautious in making sweeping conclusions due to the small size of corpora in our study, our results from gain ratio and cross entropy indicate that counts from the features of both systems will be more useful in the MUC6 corpus than in the biology corpus. We can also see that while the coverage is limited, surface words play a leading role for both systems. Gain ratio statistics for surface words in the two domains were far closer than for any other type of feature, and given that this is also the dominant knowledge type this seems to be one likely reason that the performance of systems is about the same in both domains.

We have presented the results of applying two supervised learning based models to the named entity task in two widely different domains and explained the performance through class-token ratios, entropy and gain ratio. Measures such as entropy and gain ratio have been found to have the best predictive power, although the features used to calculate gain ratio are not sufficient to describe all the information that is necessary for the named entity task. In future work we intend to extend our study to new and larger NE corpora in various domains and to try to reduce the error factor in our calculations that is a result of corpus size.

References

A. Bairoch and R. Apweiler. 1997. The SWISS-PROT protein sequence data bank and its new supplement TEmBL. *Nucleic Acids Research*, 25:31-36.

D. Bikel, S. Miller, R. Schwartz, and R. Westchapel. 1997. Nymble: a high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194-201.

A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition. In *Proceedings of the Workshop on Very Large Corpora* (WVLG'98).

L. Breiman, R. Friedman, A. Olshen, and C. Stone. 1984. *Classification and regression*
trees. Belmont CA: Wadsworth International Group.

S. Chen and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. 34th Annual Meeting of the Association of Computational Linguistics, California, USA, 24-27 June.

N. Chinchor. 1995. MUC-5 evaluation metrics. In Proceedings of the Fifth Message Understanding Conference (MUC-5), Baltimore, Maryland, USA, pages 69-78.

N. Collier, C. Nobata, and J. Tsujii. 2000. Extracting the names of genes and gene products with a hidden Markov model. In Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000), Saarbrucken, Germany, July 31st--August 4th.

M. Craven and J. Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology (ISMB-99), Heidelberg, Germany, August 6–10.

K. Fukuda, T. Tsumoda, A. Tamura, and T. Takagi. 1998. Toward information extraction: identifying protein names from biological papers. In Proceedings of the Pacific Symposium on Biocomputing'98 (PSB'98), January.

MEDLINE. 1999. The PubMed database can be found at: http://www.ncbi.nlm.nih.gov/PubMed/.

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

C. Nobata, N. Collier, and J. Tsujii. 1999. Automatic term identification and classification in biology texts. In Proceedings of the Natural Language Pacific Rim Symposium (NLPRS'2000), November.

D. Palmer and D. Day. 1997. A statistical profile of the named entity task. In Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP'97), Washington D.C., USA, 31 March – 3 April.

J.R. Quinlan. 1990. Introduction to Decision Trees. In J.W. Shavlik and T.G. Dietterich, editors, Readings in Machine Learning. Morgan Kaufmann Publishers, Inc., San Mateo, California.

J.R. Quinlan. 1993. C4.5 Programs for Machine Learning. Morgan Kaufmann Publishers, Inc., San Mateo, California.