A Statistical Profile of the Named Entity Task

David D. Palmer and David S. Day
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA
{palmer,day}@mitre.org

Abstract

In this paper we present a statistical profile of the Named Entity task, a specific information extraction task for which corpora in several languages are available. Using the results of the statistical analysis, we propose an algorithm for lower bound estimation for Named Entity corpora and discuss the significance of the cross-lingual comparisons provided by the analysis.

1 The Named Entity task

There is currently much interest, in both research and commercial arenas, in natural language processing systems which can perform multilingual **information extraction** (IE), the task of automatically identifying the various aspects of a text that are of interest to specific users. An example of IE is the **Named Entity** (NE) task, which has become established as the important first step in many other IE tasks, providing information useful for coreference and template filling. Named Entity evaluation began as a part of recent Message Understanding Conferences (MUC), whose objective was to standardize the evaluation of IE tasks (Sundheim, 1995b). Several organized evaluations have been held to determine the state-of-the-art in NE systems, and there are commercial systems available.

The goal of the NE task is to automatically identify the boundaries of a variety of phrases in a raw text, and then to categorize the phrases identified. There are three categories of named-entities defined by the guidelines: TIMEX, NUMEX, and ENAMEX. TIMEX phrases are temporal expressions, which are subdivided into date expressions (*April 7*) and time expressions (*noon EST*). NUMEX phrases are numeric expressions, which are subdivided into percent expressions (*3.2%*) and money expressions (*$180 million*). ENAMEX phrases are proper names, representing references in a text to persons (*Jeffrey H. Birnbaum*), locations (*New York*), and organizations (*Northwest Airlines*).

Evaluation of system performance for the NE task is done using an automatic scoring program (Chinchor, 1995), with the scores based on two measures - **recall** and **precision**. Recall is the percent of the "correct" named-entities that the system identifies; precision is the percent of the phrases that the system identifies that are actually correct NE phrases. The component recall and precision scores are then used to calculate a balanced *F-measure* (Rijsbergen, 1979), where $F = 2PR/(P + R)$.

Human performance on the NE task has been determined to be quite high, with F-measures better than 96% (Sundheim, 1995b). Despite the fact that some systems in recent evaluations have performance approaching this human performance, it is important to note that named-entity recognition is by no means a "solved problem." The fact that existing systems perform extremely well on mixed-case English newswire corpora is certainly related to the years of research (and organized evaluations) on this specific task in this language. Although performance by MUC-6 and MET systems is encouraging, it is not clear what resources are required to adapt systems to new languages. It is also unknown how the existing high-scoring systems would perform on less well-behaved texts, such as single-case texts, non-newswire texts, or texts obtained via optical character recognition (OCR).

There has been little discussion of the linguistic significance of performing NE recognition, or of how much linguistic knowledge is required to perform well on such an evaluation. However, any given language task should be examined carefully to establish a baseline of performance which should be attainable by any system; only then can we adequately determine the significance of the results reported on that task. In this paper we give the results of an analysis of NE corpora in six languages from the point of view of a system with no knowledge of the languages; that is, we performed an analysis based purely on the strings of characters composing the texts and the named-entity phrases. The performance of such a straw-man system, which did not use language-specific lexicons or word lists or even information about tokenization/segmentation or part-of-speech, can serve as a baseline score for comparison of more sophisticated systems.

2 The Corpora

The definition of the NE task we discuss in this paper was taken from the guidelines for the Sixth Message Understanding Conferences (MUC-6) (Sundheim, 1995a) and the recent Multilingual Entity Task (MET, May 1996), both sponsored by the TIPSTER program. MUC-6 evaluated English NE systems, and MET evaluated Spanish, Japanese, and Chinese NE systems. The Spanish, Japanese, and Chinese corpora we analyzed each consisted of the MET training documents; similarly, the English corpus contains 60 Wall Street Journal articles prepared for the MUC-6 dry-run and official evaluation. In addition to the four corpora available from the recent organized NE evaluations, we analyzed similar-sized French and

Portuguese corpora[1] which were prepared according to the MET guidelines. Table 1 shows the sources for the corpora.

Language	News service	Country
Chinese	Xinhua	China
English	Wall Street Journal	USA
French	Le Monde	France
Japanese	Kyodo	Japan
Portuguese	Radiobras	Brazil
Spanish	Agence France Presse	France

Table 1: Corpora sources.

All six corpora consisted of a collection of newswire articles, and none of the articles in any language was a translation of an article in another language. There were important differences in the makeup of these individual corpora that affected this analysis. The French corpus, for example, contained a wide range of articles from a single issue of *Le Monde*, so the topics of the articles ranged from world politics to the Paris fashion scene. The articles in the English and Spanish corpora were specifically selected (by the MUC-6 and MET evaluation organizers) because they contained references to press conferences. While the content was more homogeneous in the English corpus, the articles were nevertheless drawn from a range of several months of the *Wall Street Journal*, so the specific topics (and constituent Named Entities) were very diverse. The Chinese *Xinhua* corpus was, in contrast, extremely homogeneous. These differences demonstrate a number of difficulties presented by corpora in different languages.

In order to estimate the complexity of the NE task, we first determined the vocabulary size of the corpora involved (i.e. "count the words"), in terms of individual lexemes of the language. For our analysis of the European-language corpora, we considered a token to be any sequence of characters delimited by white space, and we ignored the case of all letters. The Japanese corpus was segmented using NEWJUMAN, the Chinese corpus with a segmenter made available by New Mexico State University. This segmentation information was used only to estimate the corpora sizes and was not used in any of the other portions of our analysis.

Since many words occurred frequently within a corpus, the linguistic *type-token distinction* was important to our analysis. An example of this distinction would be the sentence *a pound costs a pound*, which has 5 lexeme tokens and 3 lexeme types. The ratio of lexeme tokens to types, which can be thought of as the average occurrence of each lexeme, is shown in Table 2 with the vocabulary sizes of the six corpora.

	Lexeme	Lexeme	Token/
Language	Tokens	Types	Type
Chinese	34782	4584	7.6
English	24797	5764	4.3
French	35997	8691	4.1
Japanese	21484	3655	5.9
Portuguese	42621	7756	5.5
Spanish	31991	7850	4.1

Table 2: Corpora size by lexeme.

Table 3 shows the total number of NE phrases for each language, as well as a breakdown of total phrases into the three individual categories.

Language	NE	TIM	NUM	ENA
Chinese	4454	17.2%	1.8%	80.9%
English	2242	10.7%	9.5%	79.8%
French	2321	18.6%	3.0%	78.4%
Japanese	2146	26.4%	4.0%	69.6%
Portuguese	3839	17.7%	12.1%	70.3%
Spanish	3579	24.6%	3.0%	72.5%

Table 3: NE phrases, by subcategory.

2.1 NUMEX and TIMEX phrases

From Table 3 we see that TIMEX and NUMEX phrases together composed only 20-30% of all NE phrases in each language. Furthermore, these phrases were the easiest to recognize, because they could be represented by very few simple patterns. Upon inspection of the corpora, for example, we were able to represent nearly all NUMEX phrases in each of the six corpora with just 5 patterns.[2] Similarly, given a simple list of the basic temporal phrase words for a language (months, days of the week, seasons, etc.), it was possible to construct a series of patterns to represent most of the TIMEX phrases.[3] We were able to represent at least 95% of all TIMEX in each language in similar ways with just a few patterns (less than 30 per language), constructed in a few hours. Since we found most NUMEX and TIMEX phrases to be easy to recognize, we therefore restricted our further analysis of the corpora to ENAMEX phrases, which proved to be significantly more complex.

2.2 ENAMEX phrases

Table 4 shows the numbers of ENAMEX phrases tokens contained by the six corpora. The average occurrence of each token in each language was quite low (much lower than the average occurrence of each lexeme), which indicated that many phrases occurred very infrequently in the corpus.

	ENAMEX	ENAMEX	Token/
Language	Tokens	Types	Type
Chinese	3605	887	4.1
English	1789	840	2.1
French	1820	1085	1.7
Japanese	1493	614	2.4
Portuguese	2698	981	2.8
Spanish	2593	1177	2.2

Table 4: Corpora size by ENAMEX phrases.

Nevertheless, a large number of all phrase tokens could be accounted for by a few frequently-occurring phrase types. For example, the Chinese corpus contained 2156 total LOCATION phrases, but 449 of these locations (20.8%) could be

[1] The French corpus was prepared by Marc Vilain; the Portuguese corpus was prepared by Sasha Caskey.

[2] An example of a NUMEX pattern representing a Spanish PERCENT would be a sequence of digits followed by either the percent sign (%) or the words "por ciento".

[3] An example of a NUMEX pattern representing a Spanish DATE would be the name of a month (or its abbreviation) followed by a sequence of digits (the day), optionally followed by a comma and another sequence of digits (the year).

accounted for by the three common Chinese words for *China*. Figure 1 shows a graph of the cumulative percentage of all phrases of the corresponding category represented by the x most frequently-occurring phrases of that type in the given language.

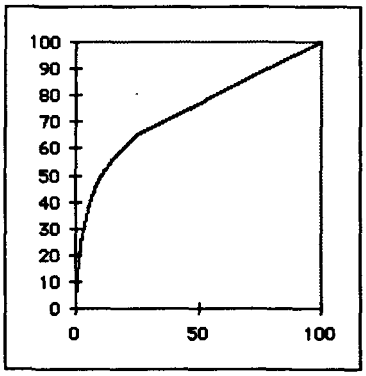

Figure 1: Graph of the cumulative % of phrase tokens provided by % of phrase types.

The graph shows a similar shape for all subcategories of ENAMEX phrases in all the languages investigated, although the rate of increase varies slightly. It is clear from the classic Zipfian distribution (cf. (Zipf, 1932; Zipf, 1949)) shown by the graph that a significant percentage of the ENAMEX phrase tokens could be represented by a small amount of frequently-occurring phrase types. However, Zipf's law also tells us that a non-trivial percentage of the phrases (those in the tail of the graph) are very infrequent, most likely never occurring in any amount of training data.

Unlike the distribution of the overall NE phrases, the relative proportion of constituent ENAMEX phrase subcategories (PERSON, LOCATION, and ORGANIZATION) varied greatly by language. The breakdown by ENAMEX phrase subcategory is shown in Table 5.

Language	Org	Loc	Pers
Chinese	20.2%	59.8%	20.0%
English	56.2%	14.5%	29.2%
French	33.8%	30.0%	38.1%
Japanese	39.2%	40.8%	20.0%
Portuguese	49.9%	19.5%	30.1%
Spanish	28.6%	43.5%	27.9%

Table 5: ENAMEX phrases by subcategory.

The significance of this result is that each ENAMEX phrase subcategory had to be treated as equivalent. It was not possible to focus on a particular subcategory to obtain a consistently high score. In other words, a strategy that focuses on locations would do well on the Chinese corpus where locations comprise 59.8% of the ENAMEX phrases, but would do poorly on the English corpus, where locations are only 14.5% of the ENAMEX.

3 Training and ambiguity

A logical question to pose is, "How well can our system perform if it simply memorizes the phrases in the training texts?" Since high performance on training texts is meaningless if a system performs poorly on new, unseen texts, we estimated the performance of a simple memorization algorithm on unseen data. For our simple system, the answer to the question depended on the **vocabulary transfer rate** of the corpus, the percentage of phrases occurring in the training corpus which also occurred in the test corpus. To measure the vocabulary transfer rate for the six corpora, we randomly divided each corpus into a training set and a test set, with each test set containing about 450 ENAMEX phrases, and each training set containing all remaining phrases. We then examined the ENAMEX phrases in the training set to determine how many also occurred in the test set.

The results of this experiment showed that, to a certain extent, a word list built from the training set provided reasonable performance. Just as some frequent phrase types comprised a large percentage of the phrase tokens within a corpus, a small number of phrase types from the training set accounted for many tokens in the test set. As shown by the transfer curve for the six languages in Figure 2, the transfer rate varied dramatically depending on the language, but the graph has the same shape for each, even though the six corpora contained different amounts of training data (thus the lines of different length).

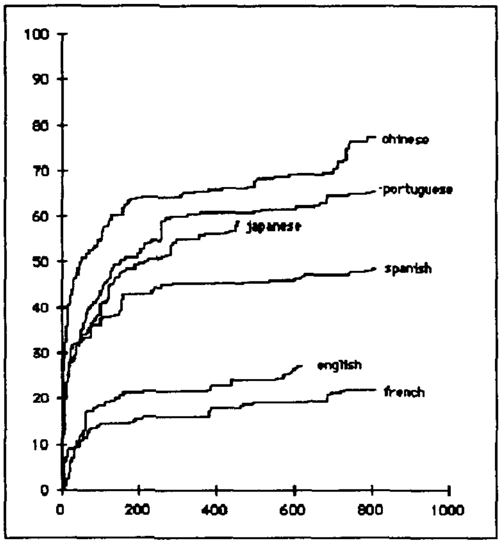

Figure 2: Graph of the cumulative test phrase tokens (%) covered by training phrase types.

In each language, the transfer rate for the most frequent phrase types (the steep part of the graph) was quite high; however, the graph rapidly peaks and leaves a large percentage of the phrases uncovered by the training phrases. The remaining "uncovered" phrases can only be recognized by means other than "memorization," such as by examining contextual clues. Table 6 shows the transfer rates of phrase tokens.

The accuracy of the pure memorization can be reduced by two forms of ambiguity. Phrases or parts of phrases can occur within two or more named-entity categories, such as the string *Boston*, which by itself is a location but within *Boston Red Sox* is an organization. In most cases this ambiguity can be resolved using a simple longest-match heuristic. Another source of ambiguity occurs when a string can occur both as a

Language	Overall ENAMEX	Org	Loc	Pers
Chinese	73.2%	46.9%	87.1%	42.6%
English	21.2%	17.7%	42.7%	13.3%
French	23.6%	13.4%	45.9%	11.2%
Japanese	59.2%	56.2%	72.7%	37.5%
Portuguese	61.3%	56.4%	57.4%	47.9%
Spanish	48.1%	49.8%	71.4%	13.7%

Table 6: Vocabulary transfer (tokens).

Language	Lower Bound
Chinese Xinhua	71.8
English WSJ	38.4
French Le Monde	34.5
Japanese Kyodo	70.1
Portuguese Radiobras	71.3
Spanish AFP	59.3

Table 7: Estimated lower bounds.

NE phrase and as a non-phrase, such as *Apple*, which would sometimes refer to the computer company (and thus be tagged an organization) and sometimes refer to the fruit (and thus not be tagged at all). Such cases, although infrequent, would result in precision errors which we do not factor into the following estimation of a recall lower bound.

4 Estimating a lower bound

Given the above statistical analysis, we estimated a baseline score for our straw-man algorithm on the NE task, a score which should easily be attainable by any system attempting to perform the task. First, we estimated that any system should be able to recognize a large percentage of NUMEX and TIMEX phrases; our experience indicates that 95% is possible due to the small number of patterns which compose most of these phrases.

In order to estimate a lower bound for ENAMEX recognition, we relied on the transfer graph in Figure 2. It is clear from the graph that the contribution of the training data has leveled off in each language by the time the number of training types is roughly equal to the size of the test data (450 in this case). Selecting this point on the graph allowed us to directly compare memorization performance for the six languages. An ideal memorization-based algorithm would be able to recognize phrases according to the transfer rate corresponding to this amount of training data. Our lower bound formula would thus be

$$((N_{NUMEX} + N_{TIMEX}) * \alpha) + (N_{ENAMEX} * T_{ENAMEX})$$

where
$\alpha = 0.95$ (in our experience)
N_{cat} = Percentage of NE phrases represented
by category (from Table 3)
T_{ENAMEX} = ENAMEX transfer rate
(from Figure 2)

The resulting lower bound scores, shown in Table 7, were surprisingly high, indicating that a very simple NE system could easily achieve a recall above 70 for some languages. The range of lower bound scores can partly be attributed to the differences in corpus makeup discussed in Section 3, but the range also illustrates the large score differences which are possible from one corpus to the next.

The upper bounds of memorization algorithms implied by the preceding analysis do not require that a deeper understanding of the linguistic phenomena of a target language is necessary to generalize NE recognition in unseen test data. Contextual clues can improve the expected score of a baseline system without requiring extensive linguistic knowledge. Just as most of the TIMEX and NUMEX phrases in any language can be recognized upon inspection using simple pattern matching, a large percentage of the ENAMEX phrases could be codified given an adequate analysis of the phrasal contexts in the training documents. Furthermore, lists of titles, geographic units, and corporate designators would assist this contextual analysis and improve the expected baseline. Indeed, such simple strategies drive most current NE systems.

5 Discussion

The results of this analysis indicate that it is possible to perform much of the task of named-entity recognition with a very simple analysis of the strings composing the NE phrases; even more is possible with an additional inspection of the common phrasal contexts. The underlying principle is Zipf's Law; due to the prevalence of very frequent phenomena, a little effort goes a long way and very high scores can be achieved directly from the training data. Yet according to the same Law that gives us that initial high score, incremental advances above the baseline can be arduous and very language specific. Such improvement can most certainly only be achieved with a certain amount of well-placed linguistic intuition.

The analysis also demonstrated the large differences in languages for the NE task, suggesting that we need to not only examine the overall score but also the ability to surpass the limitations of word lists, especially since extensive lists are available in very few languages. It is particularly important to evaluate system performance beyond a lower bound, such as that proposed in Section 4. Since the baseline scores will differ for different languages and corpora, scores for different corpora that appear equal may not necessarily be comparable.

References

Nancy Chinchor. 1995. MUC-5 evaluation metrics. In *Proceedings of the Fifth Message Understanding Conference (MUC5)*, pages 69–78, Baltimore, Maryland.

Steven Maiorano and Terry Wilson. 1996. Multilingual Entity Task (MET): Japanese Results. In *Proceedings of TIPSTER Text Program (Phase II)*, May.

Roberta Merchant and Mary Ellen Okurowski. 1996. The Multilingual Entity Task (MET) Overview. In *Proceedings of TIPSTER Text Program (Phase II)*, May.

MET. May 1996. Task definition. Multilingual Entity Task.

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

Beth Sundheim. 1995a. MUC6 named entity task definition, Version 2.1. In *Proceedings of the Sixth Message Understanding Conference (MUC6)*.

Beth M. Sundheim. 1995b. Overview of results of the MUC-6 evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC6)*.

G. Zipf. 1932. *Selected Studies of the Principle of Relative Frequency in Language*. Harvard University Press, Cambridge, MA.

G. Zipf. 1949. *Human Behavior and the principle of least effort*. Hafner, New York.