Named Entity Extraction from Noisy Input: Speech and OCR

David Miller, Sean Boisen, Richard Schwartz, Rebecca Stone, Ralph Weischedel
BBN Technologies
70 Fawcett Street
Cambridge, MA 02138
dmiller@bbn.com, boisen@bbn.com, schwartz@bbn.com, rwstone@bbn.com, weischedel@bbn.com

Abstract

In this paper, we analyze the performance of name finding in the context of a variety of automatic speech recognition (ASR) systems and in the context of one optical character recognition (OCR) system. We explore the effects of word error rate from ASR and OCR, performance as a function of the amount of training data, and for speech, the effect of out-of-vocabulary errors and the loss of punctuation and mixed case

1 Introduction

Information extraction systems have traditionally been evaluated on online text with relatively few errors in the input. For example, this description of the Nominator system (Wacholder et al. 1997) would apply to several other systems: "We chose The Wall Street Journal corpus because it follows standard stylistic conventions, especially capitalization, which is essential for Nominator to work." The real-world challenge, however, is pointed out in Palmer and Day (1997): "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 text obtained via optical character recognition (OCR)."

In this paper we explore how performance degrades on noisy input, in particular on broadcast news (speech) and on newspaper (printed matter). Error rates of automatic speech recognizers (ASR) on broadcast news are still very high, e.g., 14-28% word error. Though character error can be very low for laser printer output, word error rates of 20% are possible for OCR systems applied to newsprint or low-quality printed matter.

In this paper, we evaluate a learning algorithm, a hidden Markov model (HMM), for named entity extraction applied to human transcripts of news, to transcripts without case or punctuation (perfect speech output), to errorful ASR output and to OCR output. Extracting information from noisy sources poses the following challenges, which are addressed in the paper.

- Since speech recognizers do not generate mixed case nor punctuation, how much do case and punctuation contribute to recognizing names in English? (Section 3.) Note that these challenges also arise in languages without case to signal proper nouns (e.g., Chinese, German, Japanese), in mono-case English or informal English (e.g., emails).
- How much will performance degrade with increasing error in the input? (Section 4.)
- How does closed vocabulary recognition affect information extraction performance? (Section 5)
- For the learning algorithm employed, how much training and effort are required? (Section 6)
- How much do lists of names contribute to performance? (Section 7)

2 Algorithms and Data

2.1 Task Definition and Data

The named entity (NE) task used for this evaluation requires the system to identify all named locations, named persons, named organizations, dates, times, monetary amounts, and percentages. The task definition is given in Chinchor, et al. (1998).

For speech recognition, roughly 175 hours of news broadcasts (roughly 1.2m words of audio) were available from the National Institute for Science and Technology (NIST) for training. All of that data includes both the audio and a manual transcription. The test set consisted of 3 hours of news (roughly 25k words).

For the combined OCR/NE system, the OCR component was trained on the University of Washington English Image Database, which is comprised primarily of technical journal articles. The NE system was trained separately on 690K words of 1993 Wall Street Journal (WSJ) data (roughly 1250 articles), including development data from the Sixth Message Understanding Conference (MUC-6) Named Entity evaluation. The test set was approximately 20K words of separate WSJ data (roughly 45 articles), also taken from the MUC-6 data set. Both test and training texts were original text (no OCR errors) in mixed case with normal punctuation. Printing the on-line text, rather than using the original newsprint, produced the images for OCR, which were all scanned at 600 DPI.

2.2 Algorithms

The information extraction system tested is IdentiFinder(TM), which has previously been detailed in Bikel et al. (1997, 1999). In that system, an HMM labels each word either with one of the desired classes (e.g., person, organization, etc.) or with the label NOT-A-NAME (to represent "none of the desired classes"). The states of the HMM fall into regions, one region for each desired class plus one for NOT-A-NAME. (See Figure 2-1.) The HMM thus has a model of each desired class and of the other text. Note that the implementation is not confined to the seven name classes used in the NE task; the particular classes to be recognized can be easily changed via a parameter.

Within each of the regions, we use a statistical bigram language model, and emit exactly one word upon entering each state. Therefore, the number of states in each of the name-class regions is equal to the vocabulary size. Additionally, there are two special states, the START-OF-SENTENCE and END-OF-SENTENCE states. In addition to generating the word, states may also generate features of that word.

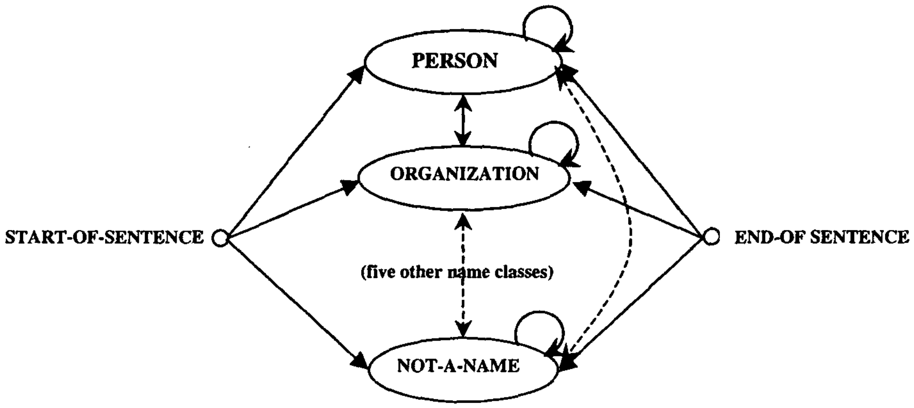

Figure 2-1: Pictorial representation of conceptual model

3 Effect of Textual Clues

The output of each of the speech recognizers is in SNOR (speech normalized orthographic representation) format, a format which is largely unpunctuated and in all capital letters (apostrophes and periods after spoken letters are preserved). When a typical NE extraction system runs on ordinary English text, it uses punctuation and capitalization as features that contribute to its decisions. In order to learn how much degradation in performance is caused by the absence of these features from SNOR format, we performed the following experiment. We took a corpus that had full punctuation and mixed case and preprocessed it to make three new versions: one with all upper case letters but punctuation preserved, one with original case but punctuation marks removed, and one with both case and punctuation removed. We then partitioned all four versions of the corpus into a training set and a held-out test set, using the same partition in all four versions, and measured IdentiFinder's performance.

The corpus we used for this experiment was the transcriptions of the second 100 hours of the Broadcast News acoustic modelling data, comprising 114 episodes. We partitioned this data to form a training set of 98 episodes (640,000 words) and a test set of 16 episodes (130,000 words). Because the test transcriptions were created by humans, they have a 0% word error rate. The results are shown in Table 3-1. The removal of case information has the greater effect, reducing performance by 2.3 points, while the loss of punctuation reduces performance by 1.4 points. The loss from removing both features is 3.4 points, less than the sum of the individual degradations. This suggests that there are some events where both mixed case and punctuation are required to lead IdentiFinder to the correct answer.

	Mixed Case	Upper Case
With punctuation	92.4	90.1
Without punctuation	91.0	89.0

Table 3-1: Effect of case and punctuation on performance(F-measure) on Broadcast News data

It should be noted that because the data are transcriptions of speech, no version of the corpus contains all the textual clues that would appear in newspaper text like the MUC-7 New York Times data. In particular, numbers are written out in words as they would be spoken, not represented using digits, and abbreviations such as "Dr.", "Jr." or "Sept." are expanded out to their full spoken word. We conclude that the degradation in performance going from newspaper text to SNOR recognizer output is at least 3.4 points in the 0% WER case, and probably more due to these other missing text clues.

4 Effect of Word Errors

4.1 Optical Character Recognition (OCR)

The OCR experiments were performed using the system described in Makhoul et al. (1998). Recognition was performed at the character level, rather than the word level, so the vocabulary is not closed (unlike the ASR results discussed in subsequent sections). Figure 4-1 shows IdentiFinder's performance under 4 conditions of varying word error rate (WER):

1. Original text (no OCR, 0% WER)
2. OCR from high-quality (laser-printed) text images (2.7% WER)
3. OCR on degraded images (13.7% WER).
4. OCR on degraded images, processed with a weak character language model (19.1% WER)

For the second and third conditions, 1.3M characters of Wall Street Journal were used for

OCR language model training: the fourth condition used a much weaker character language model, which accounts for the poorer performance.[1]

The interpolated line has been fit to the performance of the OCR-based systems, with a slope indicating 0.6 points of F-measure lost for each percentage point increase in word error. The line has been extrapolated to 0% WER: the actual 0% WER condition is 95.4, which only slightly exceeds the projected value.

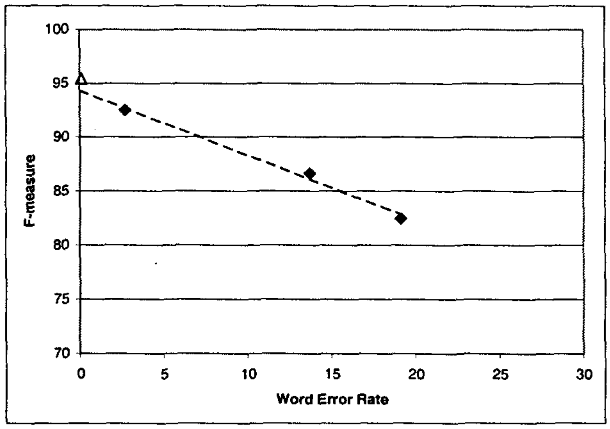

Figure 4-1: IdentiFinder Named Entity performance as a function of OCR word error rate

4.2 Automatic Speech Recognition (ASR)

Figure 5-1 shows IdentiFinder's performance on all speech systems in the 1998 Hub-4 evaluations (Przybocki, et al., 1999). These experiments were run in co-operation with NIST. The interpolated line has been fit to the errorful transcripts, and then extrapolated out to 0% WER speech. As can be seen, the line fits the data extremely well, and has a slope of 0.7 points of F-measure lost for each additional 1% of word error rate. The fact that the extrapolated line slightly overestimates the actual performance at 0% WER (given by a Δ) indicates that the degradation may be sub-linear in the range 0-15% WER.

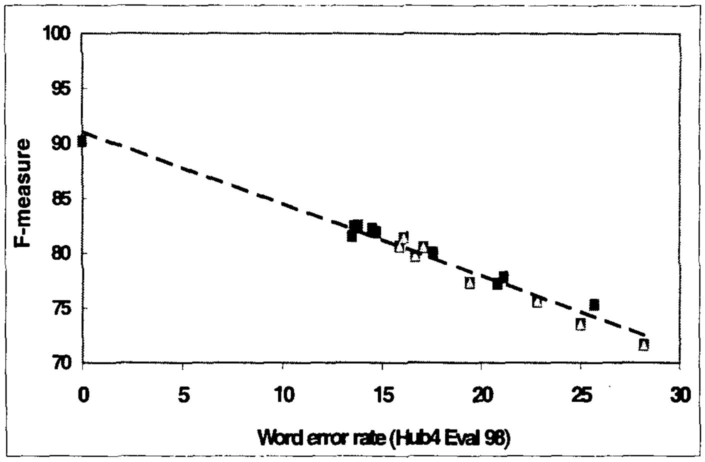

Figure 4-2: IdentiFinder named-entity performance as a function of word error rate (in cooperation with NIST)

5 Out of Vocabulary Rates for Names

It is generally agreed that out-of-vocabulary (OOV) words do not have a major impact on the word error rate achieved by large vocabulary speech recognizers doing transcription. The reason is that speech lexicons are designed to include the most frequent words, thus ensuring that OOV words will represent only a small fraction of the words in any test set. However, we have seen that the OOV rate for words that are part of named-entities can be as much as a factor of ten greater than the baseline OOV for non-name words. This could make OOV a major problem for NE extraction from speech.

To explore this, we measured the percentage of names in the Broadcast News data that contain at least one OOV word as a function of lexicon size. For this purpose, we built lexicons simply by ordering the words of the 1998 Hub-4 Language Modeling data according to

[1] These figures do not reflect the best possible performance of the OCR system: for example, when testing on degraded data, it would be usual to include representative data in training. This was not a concern for this experiment, however, which focussed on name finding performance.

Name Category	Lexicon Size							
	5K	10K	20K	40K	60K	80K	100K	120K
PERSON	34.7	52.7	69.9	85.1	89.4	91.1	91.9	93.9
ORGANIZATION	73.2	90.2	94.2	97.5	98.2	98.5	98.7	98.8
LOCATION	76.6	87.1	92.2	96.2	97.5	98.0	98.8	99.1
TIME	97.0	97.0	99.0	100	100	100	100	100
MONEY	94.4	98.2	98.8	100	100	100	100	100
DATE	96.1	99.3	99.8	100	100	100	100	100
PERCENT	98.9	99.3	100	100	100	100	100	100

Table 5-1: Percentage of in-vocabulary events as a function of lexicon size.

frequency, and truncating the list at various lengths. The percentage of in-vocabulary events of each type as a function of lexicon size is shown in Table 5-1.

Most modern speech recognizers employ a vocabulary of roughly 60,000 words; using a larger lexicon introduces more errors from acoustic perplexity than it fixes through enlarged vocabulary. It is clear from the table that the only name category that might suffer a significant OOV problem with a 60K vocabulary is PERSONs. One might imagine that a more carefully constructed lexicon could reduce the OOV rate for PERSONs while still staying within the 60,000 word limit. However, even if a cleverly designed 60K lexicon succeeded in having the name coverage of the frequency-ordered 120K word lexicon (which contains roughly 40,000 more proper names than the 60K lexicon), it would reduce the PERSON OOV rate by only 4% absolute.

Given that PERSONs account for roughly 50% of the named-entities in broadcast news, the maximum gain in F measure available for doubling the lexicon size is 2 points. Moreover, this gain would require that every PERSON name added to the vocabulary be recognized properly -- an unlikely prospect, since most of these words will not appear in the acoustic training for the recognizer. For these reasons, we conclude that the OOV problem is not a major factor in determining NE performance from speech.

6 Effect of training set size

6.1 Automatic Speech Recognition

We have measured NE performance in the context of speech as a function of training set size and found that the performance increases logarithmically with the amount of training data for 15% WER test data as well as for 0% WER input. However the growth rate is slower for 15% WER test data. We constructed small training sets of various size by randomly selecting sets of 6, 12, 25, and 49 episodes from the second 100 hours of annotated Broadcast News training data. We also defined a training set of 98 episodes from the second 100 hours, as well as sets containing the full 98 episodes plus some or all of the first 100 hours of Broadcast News training. Our largest training set contained 1.2 million words, and our smallest a mere 30,000 words. All training data were converted to SNOR format.

For each training set, we trained a separate IdentiFinder model and evaluated it on two versions of the 1998 Hub4-IE data -- the 0% WER transcription created by a human, and an ASR transcript with 15%. The results are plotted in Figure 6-1. The slopes of the interpolated lines predict that IdentiFinder's performance on 15% WER speech will increase by 1.5 points for each additional doubling of the training data, while performance goes up 1.8 points per doubling of the training for perfect speech input.

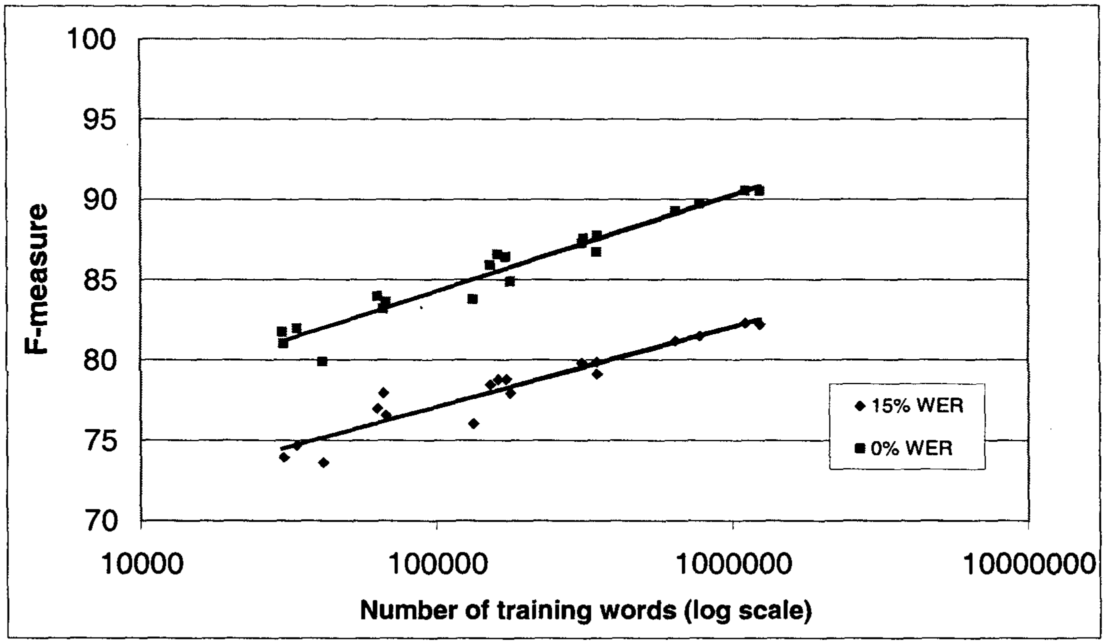

Figure 6-1: Performance as a function of training data for speech.

Possibly, the difference in slope of the two lines is that the real value of increasing the training set lies in increasing the number of distinct rare names that appear. Once an example is in the training, IdentiFinder is able to extract it and use it in test. However, when the test data is recognizer output, the rare names are less likely to appear in the test, either because they don't appear in the speech lexicon or they are poorly trained in the speech model and misrecognized. If they don't appear in the test, IdentiFinder can't make full use of the additional training, and thus performance on errorful input increases more slowly than it does on error-free input text.

6.2 Optical Character Recognition

A similar relationship between training size and performance is seen for the OCR test condition.

The training was partitioned by documents into equal sized sets:

Partition size	Training Size
Eighth	77.5 K words
Quarter	155 K words
Half	310 K words
Whole	620 K words

Using the same test set, each partition was used to train a separate model, which was then evaluated on the different word error conditions: performance was then averaged across each partition size to produce the data points below.

Input Word Error Rate (WER)	Eighth	Quarter	Half	Whole
0% WER (Original text)	92.4	93.7	94.3	95.3
2.7% WER	90.0	90.8	91.6	92.5
13.7% WER	84.3	85.2	86.0	86.6
19.1% WER	79.6	80.4	80.8	82.5

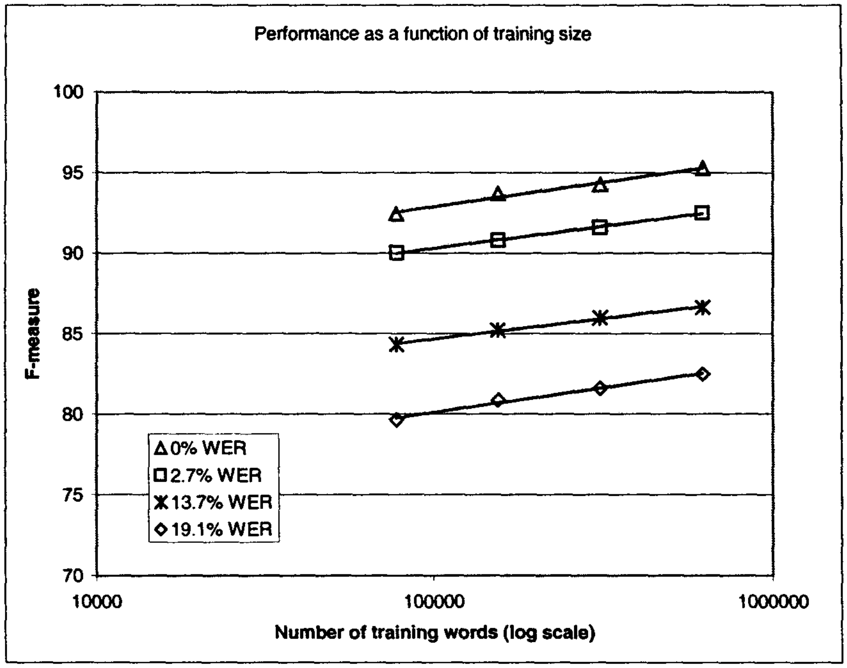

Figure 6–2: Performance as a function of training data for OCR.

While this graph of this data in Figure 6-2 shows a logarithmic improvement, as with the ASR experiments, the rate of improvement is substantially less, roughly 0.9 increase in F-measure for doubling the training data. This may be explained by the difference in difficulty between the two tests: even with only 77.5k words of training, the 0% WER performance exceeds the ASR system trained on 1.2M words.

7 Effect of Lists

Like most NE extraction systems, IdentiFinder can use lists of strings of known to be names to estimate the probability that a word will be a name, given that it appears on a particular list. We trained two models on 1.2 million words of SNOR data, one with lists and one without. We tested on the human transcription (0% WER) and the ASR (15% WER) versions of the 1998 evaluation transcripts. Table 7-1 shows the results. We see that on human constructed transcripts, lists improve the performance by a full point, while on recognizer produced output, performance goes up by only 0.3 points.

	0% WER	15% WER
Without lists	89.5	81.9
With lists	90.5	82.2

Table 7-1: Effect of lists in the presence of speech errors.

8 Related Work and Future Work

To our knowledge, no other information extraction technology has been applied to OCR material.

For audio materials, three related efforts were benchmarked on NE extraction from broadcast news. Palmer, et al. (1999) employs an HMM very similar to that reported for IdentifFinder (Bikel et al., 1997,1999). Renals et al. (1999) reports on a rule-based system and an HMM integrated with a speech recognizer. Appelt and Martin (1999) report on the TEXTPRO system, which recognises names using manually written finite state sales.

Of these, the Palmer system and TEXTPRO report results on five different word error rates. Both degrade linearly, about .7F, with each 1% increase in WER from ASR. None report the effect of training set size, capitalization, punctuation, or out-of-vocabulary items.

Of the four systems, IdentiFinder represents state-of-the-art performance. Of all the systems evaluated, those with the simple architecture of ASR followed by information extraction performed markedly better than the system where extraction was more integrated with ASR.

In general, these results compare favorably with results reported in the Message Understanding Conference (Chinchor, et al., 1998). The highest NE score in MUC-7 was 93.39; for 0% WER, our best score was 90.5 without case and punctuation which costs about 3.4 points.

9 Conclusions

First and foremost, the hidden Markov model is quite robust in the face of errorful input. Performance on both speech and OCR input degrades linearly as a function of word error. Even, without case information or punctuation in the input, the performance on the broadcast news task is above 90%, with only a 3.4 point degradation in performance due to missing textual clues. Performance even with 15% word error degrades by only about 8 points of F for both OCR and ASR systems.

Second, because annotation can be performed quickly and inexpensively by non-experts, training-based systems like IdentiFinder offer a powerful advantage in moving to new languages and new domains. In our experience, annotation of English typically proceeds at 5k words per hour or more. This means interesting performance can be achieved with as little as 20 hours of student annotation (i.e., at least 100k words). Increasing training continually improves performance, generally as the logarithm of the training set size. On transcribed speech, performance is already good (89.3 on 0% WER) with only 100 hours or 643K words of training data.

Third, though errors due to words out of the vocabulary of the speech recognizer are a problem, they represent only about 15% of the errors made by the combined speech recognition and named entity system.

Fourth, we used exactly the same training data, modeling, and search algorithm for errorful input as we do for error-free input. For OCR, we trained on correct newswire once only for both correct text input 0% (WER) and for a variety of errorful text input conditions. For speech, we simply transformed text training data into SNOR format and retrained. Using this approach, the only cost of handling errorful input from OCR or ASR was a small amount of computing time. There were no rules to rewrite, no lists to change, and no vocabulary adjustments. Even so, the degradation in performance on errorful input is no worse than the word error rate of the OCR/ASR system.

Acknowledgments

The work reported here was supported in part by the Defense Advanced Research Projects Agency. Technical agent for part of this work was NRaD under contract number N66001-97-D-8501. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the United States Government.

References

D. E. Appelt, D. Martin, "Named Entity Extraction from Speech: Approach and Results Using the TextPro System," *Proceedings Of The DARPA Broadcast News Workshop, February 28-March 3,* Morgan Kaufmann Publishers, pp 51-54 (1999).

D. Bikel, S. Miller, R. Schwartz, and R. Weischedel, 'Nymble: a High-Performance Learning Name-finder". In *Fifth Conference on Applied Natural Language Processing,* (published by ACL) pp 194-201 (1997).

D. Bikel, R. Schwartz, and R. Weischedel, "An Algorithm that Learns What's in a Name," *Machine Learning 34*, pp 211-231, (1999).

N. Chinchor, "MUC-7 Named Entity Task Definition Version 3.5". Available by ftp from ftp.muc.saic.com/pub/MUC/MUC7-guidelines. (1997).

N. Chincor, P. Robinson, E. Brown, "HUB-4 Named Entity Task Definition Version 4.8". Available by ftp from www.nist.gov/speech/hub4_98. (1998).

J. Makhoul, R. Schwartz, C. Lapre, and I. Bazzi, "A Script-Independent Methodology for Optical Character Recognition,", *Pattern Recognition,* pp 1285-1294 (1998).

Z. Lu, R. Schwartz, P. Natarajan, I. Bazzi, J. Makhoul, "Advances in the BBN BYBLOS OCR System," *Proceedings of the International Conference on Document Analysis and Recognition*, (1999).

D. D. Palmer, J. D. Burger, M. Ostendorf, "Information Extraction from Broadcast News Speech Data," *Proceedings Of The DARPA Broadcast News Workshop, February 28-March 3,* Morgan Kaufmann Publishers, pp 41-46 (1999).

M. A. Przybocki, J. G. Fiscus, J. S. Garofolo, D. S. Pallett, "1998 Hub-4 Information Extraction Evaluation," *Proceedings Of The DARPA Broadcast News Workshop, February 28-March 3,* Morgan Kaufmann Publishers, pp 13-18 (1999).

S. Renals, Y. Gotoh, R. Gaizauskas, M. Stevenson, "Baseline IE-NE Experiments Using the SPRACH/LASIE System," *Proceedings Of The DARPA Broadcast News Workshop, February 28-March 3,* Morgan Kaufmann Publishers, pp 47-50 (1999).