

Named Entity Extraction from Broadcast News

David Miller, Richard Schwartz, Ralph Weischedel, Rebecca Stone

BBN Technologies
70 Fawcett Street
Cambridge, MA 02138

ABSTRACT

In this paper, we contrast the two tasks of named entity extraction from speech and text both qualitatively and quantitatively in the context of the DARPA 1998 Hub4e-IE evaluation. We will present some top level observations and a detailed engineering analysis of our system’s failures and successes. We explore the effects of word error rate, loss of textual clues, amount of training data, changes in guidelines, and out-of-vocabulary errors.

1. Introduction

BBN used the IdentiFinder(tm) system (described in [1]) to perform the Named-Entity extraction spoke of the 1998 Hub4 DARPA evaluation. We annotated with named-entity markup the 175 hours of Broadcast News acoustic modelling data, and trained IdentiFinder’s statistical models on it. In test, IdentiFinder segmented the input text into paragraphs by splitting at story boundaries (when working from text) or at 1 second silences (when working from speech).

Table 1 shows our official evaluation results for all five transcript conditions. The word error rates (WER) of these transcripts vary from 0% to 28.3%, and in all cases the transcript uses SNOR input (all upper case text with minimal punctuation).

We offer three top level observations about these results.

First, the scores on this evaluation set are very good. The 90.6 achieved on 0% WER speech output is the same performance achieved on the MUC-7 New York Times data. This is a pleasant surprise, since the NYT data was mixed case, had complete and consistent punctuation, and used digit strings to represent many dates, times, and dollar amounts. The absence of these clues should make the 0% WER speech problem substantially harder. Indeed, we show in Section 3 that restoring case and punctuation information to Broadcast News data raises performance by about 3.4% absolute. One possible reason for the higher accuracy is that the change in annotation guidelines from MUC-7 to Hub4-IE has made the task easier, but measurements (see Section 4) show this not to be the

Transcript	WER	F
Reference	0.0%	90.6
Baseline1	13.5%	81.5
Baseline2	14.5%	82.6
Baseline3	28.3%	70.3
BBN1	14.7%	82.2

Table 1: F-measures for BBN’s tagger on various eval98 transcripts.

case. Another possibility is that this test set is intrinsically easier, or is better matched to the training than in the MUC-7 case.

Second, we used exactly the same training data, modeling, and decoding processing for the 0% WER speech as for the errorful speech. The only adjustment we made to use the system on speech test data was to reprocess the training data into SNOR format. Using this approach there were no rules to rewrite, no lists to change, and no vocabulary adjustments. Even so, the degradation in performance on speech output is substantially less than the speech WER. This approach is thus effective and inexpensive.

Third, to adjust our system to the domain of Broadcast News, we annotated 175 hours of training data. The annotation was performed by college students without any specialized knowledge in computational linguistics. Because annotation can be performed quickly and inexpensively by non-experts, training-based systems like IdentiFinder hold a powerful advantage in moving to new languages and new domains. We estimate that the annotation process took one person-month of annotator time per 500,00 words, including double annotation, adjudication, test-on-train cycles, and supervision. The first 100 hours of transcribed data required less than two person months, and was made available to the community and used by all sites in this evaluation. The remaining 75 hours of data were not ready in time for general distribution, but will be made available as soon as possible.

The remainder of this paper presents experiments supporting the comments above, and explores many of the points more thoroughly. In particular, we focus on the effects of word error rate, loss of textual clues, out-of-vocabulary errors, amount of training data, and incorporating prior knowledge.

2. Effect of WER

The evaluation results presented in Table 1 are plotted in Figure 1. While these data are too scant to make any firm conclusions, it appears that the F-measure degrades linearly with increased word error rate, with a slope of 0.7 points of F-measure lost per 1% of additional word error rate. This hypothesis is substantially bolstered by the results NIST has reported ([4]) when using IdentiFinder (trained on different data) as the reference tagger for the 1998 Hub4 evaluation. Figure 2 shows IdentiFinder’s performance on all the primary and 10X-spoke systems for this evaluation. 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 the same 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 must be sub-linear in the range

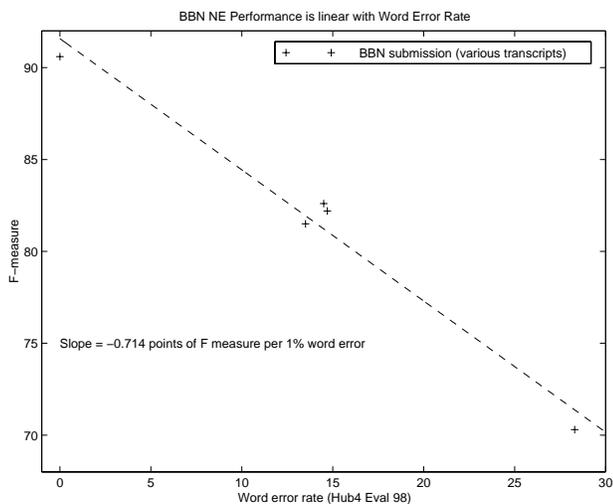


Figure 1: BBN named-entity performance as a function of word error rate

0-15% WER.

That IdentiFinder’s performance degrades (at worst) linearly with a slope of 0.7 implies that not every speech error causes a named-entity error. This is in some way a measure of IdentiFinder’s ability to label a word string correctly despite the fact that some of the words in that string are mis-transcribed. It is IdentiFinder’s ability to use word contexts that makes this compensation possible, and it is IdentiFinder’s reliance on local contexts only that prevents the effect of an incorrectly transcribed word from propogating out to the labelling of the full sentence.

3. Effect of Textual Clues

The output of the Byblos speech recognizer is in SNOR format, a format which is largely unpunctuated and in all capital letters (apostrophes and periods after spoken letters are preserved). When IdentiFinder runs on ordinary text, it uses punctutaion and capitalization as features that contribute to its decisions. In order to learn how much degradation in performance was 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 pre-processed it to make three new versions: one with all upper case letters but punctuation preserved, one with original casing 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 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 2. The removal of case information has the greater effect, reducing performance by 2.3 points, while the loss of punctuation reduces per-

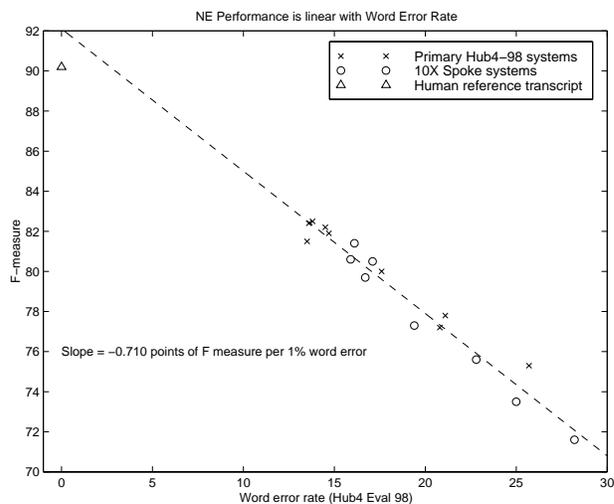


Figure 2: IdentiFinder named-entity performance as a function of word error rate (as reported by NIST)

formance 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.

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 NYT 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 New Annotation Guidelines

The annotation guidelines for the 1998 Hub4-IE spoke differ somewhat from those used in the MUC-6 and MUC-7 evaluations ([2, 3]). The major changes involve the treatment of relative times (e.g. “today”, “last month”) and the treatment of artifacts (e.g. when is “New York Times” an organization and when not?). Because we had previously annotated the second 100 hours of the Broadcast News data following the MUC-7 guidelines, we were able to measure the change in performance caused by the change in annotation guidelines. We did not have a version of the 1998 evaluation transcripts

	Mixed case	Upper case
with punctuation	92.4	90.1
without punctuation	91.0	89.0

Table 2: Effect of case and punctuation on performance. F-measure on 130K words of held-out Broadcast News data.

guidelines	F-measure
MUC-7	89.26
1998 HUB4-IE	89.03

Table 3: Performance under different annotation guidelines. Test set held-out from second 100 hours of training.

annotated by the old rules, so we instead tested on a held out portion of the training, following the partition described in Section 3. We preprocessed the data to conform to SNOR format rules, stripping punctuation and upcasing all the letters. The results are thus comparable to 0% WER speech data.

We trained two models, one with the data annotated according to the MUC-7 rules, and one with the data annotated according to the Hub4-IE rules. We tested with each model, and scored each result using a reference key annotated with the matching rules. As Table 3 shows, performance on the task was essentially the same – only 0.2% harder under the new rules.

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 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 4.

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, 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

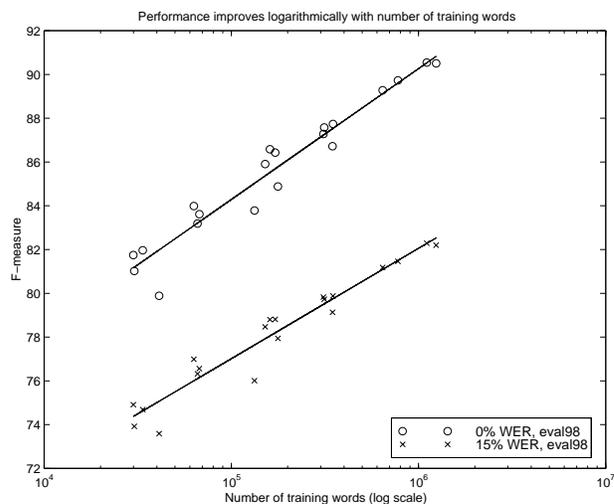


Figure 3: Performance as a function of training data. Eval 1998 “known reference” and BBN primary submission test sets.

performance from speech.

6. Effect of training set size

We have previously reported that for text, performance goes up roughly logarithmically with the amount of training data ([1]). We have remeasured this in the context of speech and found that the trend holds for 15% WER test data as well as for 0% WER input, but with a different constant as the growth rate. 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 the BBN Byblos-produced 15% WER transcript. The results are plotted in Figure 3. 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.¹

One possible explanation for 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

¹In truth, we are skeptical that performance would continue to improve at these rates for training sets larger than 1.2 million words. Though it is an expensive proposition to test, the improvements for increased training at the high end of the curves seem to be flattening out some.

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 4: Percentage of in-vocabulary events as a function of lexicon size.

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.

7. Effect of Lists

IdentiFinder can incorporate prior knowledge into its model through lists of strings that are known to be names at some times. It does not use these lists in iron-clad rules, but rather estimates from training the probability that a word will be a name, given that appears on a particular list [1]. For example, the word "HOPE" appears in the list of locations (e.g. Hope, Arkansas), but the word is not always used in training text as a location. In general, though, words on the location list tend to be used in location contexts far more than words not on the list.

We investigated the benefit of using lists in *IdentiFinder* (which can also run without any lists) on speech output. We trained two models on 1.2 million words of SNOR data, one with lists and one without. We tested on the known reference (0% WER) and the BBN Byblos (15% WER) versions of the 1998 evaluation transcripts. Table 5 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.

The reason for the difference in improvement is likely to be that the lists offer help with rare names that *IdentiFinder* may not have seen in the training. These names appear in the error-free text, and so the full boost from lists is realized. In the recognizer output, however, many of the rare names have already been lost to recognizer error, and so the lists provide a smaller boost. This argument is similar to that used to explain why additional training data improves performance more quickly in the 0% WER setting than it does in the errorful input setting.

	0% WER	15% WER
w/o lists	89.5	81.9
with lists	90.5	82.2

Table 5: Effect of lists in the presence of speech errors. "Known reference" and BBN Byblos Eval 98 transcripts.

8. Conclusions

First and foremost, the hidden Markov model is quite robust in the face of errorful input. Performance on transcription with no errors is above 90% even without case information or punctuation in the input. Lack of punctuation and case information seems to cause only a 3.4 point degradation in performance. Performance even with 15% word error degrades by only 8%.

Second, though performance improves as the logarithm of the training set size, performance is already good (89.3 on 0% WER) with only 100 hours or 643K words of training data. The annotation was performed by college students without any specialized knowledge in computational linguistics. Because annotation can be performed quickly and inexpensively by non-experts, training-based systems like *IdentiFinder* hold a powerful advantage in moving to new languages and new domains.

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 0% WER speech as we do for the errorful speech. We simply transformed text training data into SNOR format and re-trained. Using this approach, the only cost of moving from text to speech 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 speech output is substantially less than the speech WER.

References

1. 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).
2. N. Chinchor, "MUC-7 Named Entity Task Definition Version 3.5". Available by ftp from ftp.muc.saic.com/pub/MUC/MUC7-guidelines. (1997).
3. N. Chinchor, P. Robinson, E. Brown, "HUB-4 Named Entity Task Definition Version 4.8". Available by ftp from www.nist.gov/speech/hub4_98. (1998).
4. NIST, "December 1998 Hub4e Broadcast News Benchmark Test Results". Available online at ftp://jaguar.ncsl.nist.gov/csr98/h4iene_98_official_scores_990107/index.htm (1999).