

# 1998 HUB-4 INFORMATION EXTRACTION EVALUATION

*Mark A. Przybocki, Jonathan G. Fiscus, John S. Garofolo, David S. Pallett*

National Institute of Standards and Technology (NIST)  
Information Technology Laboratory (ITL)  
Room A216 Building 225 (Technology)  
Gaithersburg, MD 20899  
E-mail: mark.przybocki@nist.gov

## ABSTRACT

This paper documents the Information Extraction Named-Entity Evaluation (IE-NE), one of the new spokes added to the DARPA-sponsored 1998 Hub-4 Broadcast News Evaluation.

This paper discusses the information extraction task as posed for the 1998 Broadcast News Evaluation. This paper reviews the evaluation metrics, the scoring process, and the test corpus that was used for the evaluation. Finally, this paper reviews the results of the first running of a Hub-4 IE-NE Evaluation.

The Baseline IE-NE evaluation, in which BBN's *IdentiFinder* was run on the primary system transcripts submitted for the Hub-4 Broadcast News evaluation, found that the transcripts generated by LIMSI's automatic speech recognition system produced the "highest" F-measure score (82.39).

In the Quasi IE-NE evaluation, where sites ran their own NE-taggers on a set of three baseline recognizer transcripts, the SRI developed tagger achieved the highest F-measure score for baseline recognizers 1 & 3, while the BBN developed tagger achieved the highest score for baseline recognizer 2.

In the Full IE-NE evaluation, where sites implemented their own NE-tagger on the their own automatic speech recognizer transcripts, BBN achieved the highest overall F-measure score of 82.22.

## 1. INTRODUCTION

The Hub 4 Broadcast News Evaluation had two new evaluation conditions in 1998. The first was a less than 10X real-time system spoke [12], and the other spoke (covered in this paper) was an "Information Extraction" (IE) spoke.

The technical objective of the IE spoke was to identify the information-carrying entities that exist in speech recognition output, and to begin to move the research focus from simple transcription toward spoken language information understanding.

The Message Understanding Conference (MUC) Community has identified information-carrying entities as important for natural language and information retrieval applications, where information is to be extracted from a news stream. [1] The

MUC community had worked for several years with entity identification in newswire text. In 1997, a pilot experiment with recognized broadcast news was conducted by MITRE and evaluated with a prototype scoring pipeline (**mscore**) which was also developed by MITRE. [2]

Following the MITRE experiment, it was decided that the creation of a common entity tagging task using broadcast news and involving the MUC community would expedite development of information extraction technologies for speech applications.

Given that the target task was to develop tagging technology for broadcast news, NIST chose to add the IE task as a spoke to its Hub-4 evaluation to capitalize on the existing infrastructure, corpora, and participant pool. NIST collaborated with MITRE and SAIC to develop the evaluation specifications, corpora, and software. The new task ultimately required the creation of a new transcription and annotation format for broadcast news. The new spoke was named "Hub-4 Information Extraction - Named Entity" (Hub-4 IE-NE).

MITRE and SAIC developed detailed guidelines for the Hub-4 IE-NE task (Hub-4 Named Entity Task Definition) [3]. NIST worked with SAIC to develop scoring software for the IE evaluation task, which involved the creation of a Recognition and Extraction Evaluation Pipeline (**reep**).

NIST selected and distributed the transcripts to be processed by research sites participating in the test, immediately following the Hub-4 Broadcast News Evaluation. After processing the data, the sites submitted their results to NIST to be scored.

## 2. EVALUATION METRICS

The Hub-4 IE-NE evaluation involved the recognition and identification of the following types of information entities in the broadcast news stream:

Named Entities:	Person, Location, Organization
Temporal Expressions:	Date, Time
Numeric Expressions:	Monetary, Percentage

Each entity is identified by placing a SGML tag around the text string that constitutes an entity tag. Detailed definitions of entity tags can be found in the document Hub 4 IE-NE Task

Definition version 4.8 [3].

The accuracy of the tagging procedure was measured by taking three measurements (content, extent and type) for each tag. "Content" defines whether or not the correct words were identified in the entity tag. "Extent" defines whether or not the correct range of words was contained in the entity tag. "Type" defines whether or not the correct category was assigned to the entity tag.

Composite evaluation metrics were formed from these three basic measurements. The scoring software developed for this evaluation automatically calculated Precision and Recall and a weighted combination of Precision and Recall called F-measure. [9]

Based on a suggestion from BBN, NIST also calculated another composite metric called Slot Error Rate [6]. Slot Error Rate is defined as the total number of slot errors divided by the total number of slots in the reference transcript.

Figure 1 shows that composite evaluation metrics, slot error rate and F-measure, correlate very strongly with word error rate. The amount of linear approximation that can be contributed to each set of points versus word error rate is shown by their corresponding R-squared values. [7]

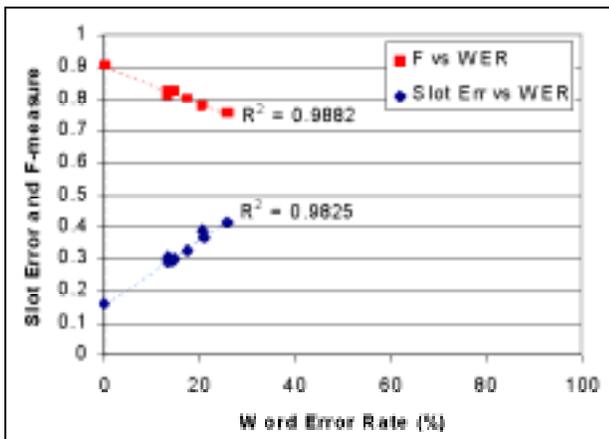


FIGURE 1: Shows a high correlation between NE-based metrics and word error rate. Note there also appears to be a strong linear correlation between F-measure and slot error rate.

### 3. EVALUATION CONDITIONS

The Hub-4 IE-NE Evaluation was designed to have three distinct tasks which involved a complex design of transcript/tagger combinations. Transcripts existed in the form of the Hub-4 Broadcast News reference key and in the form of various automatic speech recognizer generated transcripts that had processed the Hub-4 test data. There were several site-developed NE-taggers, one of which (BBN's Identifinder) NIST used as the baseline tagger.

The Hub-4 IE-NE tasks were as follows:

### 3.1 Baseline IE-NE Evaluation

The purpose of the Baseline IE-NE Evaluation was to explore the use of IE-NE based metrics (such as F-measure and slot error rate) as alternative metrics for recognition performance. In the Baseline IE-NE Evaluation, NIST processed the primary system submission files of the Hub-4 participants with the "Baseline Tagger".

NIST used a version of BBN's Identifinder as the baseline tagger. NIST trained Identifinder on the second 100 hours of broadcast news training data which included approximately 33-hours of data tagged by MITRE, and approximately 67 hours of data tagged by BBN.

In order to calibrate the baseline tagger's performance, NIST ran Identifinder on the reference transcripts of the 1998 development data supplied with the SAIC-developed software package IEEVAL. The following F-measure scores were obtained: Overall=88, Content=91, Extent=84, and Type=89. (These numbers are provided to aid researchers in comparing the performance of their NE-tagger (possibly trained on the same training data) to the tagger referred to in this paper as the *baseline tagger*.)

### 3.2 Quasi IE-NE Evaluation

The purpose of the Quasi IE-NE Evaluation was to perform tagger comparisons. In the Quasi IE-NE Evaluation participants were to implement their own NE-taggers on a set of transcripts from three baseline recognizers and submit their results to NIST for scoring. For comparison purposes, each site also ran their tagger on the reference transcript (0.0% WER).

The three recognizers that NIST selected for use as baseline ASR transcripts were as follows: IBM's primary system from the Hub-4 evaluation (which ran at a WER of 13.5%). Dragon's primary system from the Hub-4 evaluation (which ran at a WER of 14.5%). NIST's copy of Sphinx-III using 1997 acoustic models and 1998 language models, (which ran at a WER of 28.3%).

### 3.3 Full IE-NE Evaluation

The purpose of the Full IE-NE Evaluation was to implement the complete information extraction paradigm. In the Full IE-NE evaluation, sites generated their own ASR transcripts and implemented their own NE-tagger. Collaboration was encouraged between sites that worked on the problem of entity tagging but not on automatic speech recognition.

Sites were to tag their own ASR transcripts, preferably the same transcripts submitted for the Hub-4 Broadcast News Evaluation, and to provide the resulting tagged file to NIST for scoring.

## 4. PARTICIPANTS

The submission files from nine research sites were

automatically processed for the Baseline IE-NE evaluation, the same nine participants that are identified in the paper that documents the Hub-4 Broadcast News Evaluations [12].

Four sites participated in the Quasi IE-NE evaluation:

- GTE Internetworking's BBN Technologies (BBN)
- Collaborative effort involving Cambridge University's Engineering Department, Sheffield University, and the International Computer Science Institute (SPRACH)
- SRI International (SRI)
- Collaboration between Boston University and MITRE corporation (baseline recognizer #3 only).

All four of these sites participated in the Full IE-NE Evaluation, with SRI submitting results on their primary system and their less than 10X real-time system. MITRE made use of SRI's ASR transcripts in order to participate in the Full IE-NE Evaluation.

For each evaluation condition SPRACH submitted two systems: a rule based system and a statistical modeled system [13].

## 5. TEST CORPUS PROPERTIES

The evaluation data used for each of the Hub-4 IE-NE evaluation conditions was the three hour Hub-4 1998 test set.

The IE-NE reference data was obtained by having human annotators from MITRE and SAIC annotate the official 1998 Hub-4 reference transcript. The annotators used MITRE's Alembic Workbench [5] to perform the IE-NE mark-up. The results from three annotators were reconciled into one official IE-NE reference transcript.

The IE-NE reference transcript contains 1,765 entity tags. As figure 2 clearly shows the ENAMEX tag is the dominant tag type in the test set. ENAMEX tags represented 88% of all annotated tags in the test set, while both TIMEX and NUMEX tags represented only 6% of the entity tags in the test set.

Further investigation showed that there was a great redundancy on entities in the test set. For example, there were 1565 ENAMEX tags in the test set, but there were only 615 ENAMEX entities (an entity is a unique tag). This is shown graphically in figure 2; note that location-tags are the most often duplicated tag type. Another point of interest is that the 11 most frequently occurring ENAMEX entities (11 of 615) accounted for 20% of all ENAMEX tags.

A final observation from figure 2 reveals that there were approximately 220 ENAMEX tags, 30 NUMEX tags, and 15 TIMEX tags, that did not appear in the 100 hours of training data.

## 6. EVALUATION SCORING

The scoring pipeline for this evaluation was very complicated. A fully functioning version of IEEVAL (ieeval0.7) [11] was

developed after a lengthy debugging process just prior to the submission of results deadline.

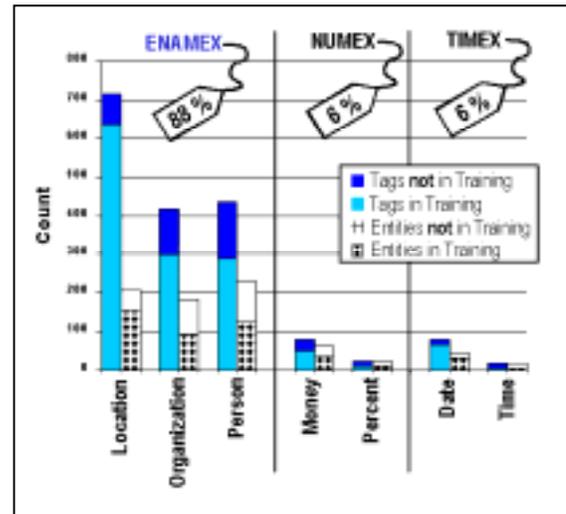


FIGURE 2: Test corpus break-down of entity-tags. Grey shaded area for tags, white and checkered bars for entities.

The scoring pipeline was modeled after MITRE's **mscore**. Where as **mscore** assumed a predefined dictionary, IEEVAL had to deal with several Hub 4 scoring difficulties, including; transcription ambiguity, lexical ambiguity, time alignment of long errorful transcripts, and word segmentation. This pipeline was a combination of three NIST packages; transcription filtering, phonetic alignment (**aldistsm**), and speech recognition scoring software (**sclite**) in conjunction with the original MUC scorer (**mscore**).

NIST, SAIC, and MITRE collaborated to build new a transcription format, Universal Transcription Format (UTF), that was able to handle these various scoring difficulties.

The evaluation design required NE-taggers to insert tags into UTF files without modifying the actual text of the files. The IEEVAL scoring software takes as input a reference and a hypothesis file, both in UTF format, and outputs scoring statistics.

## 7. EVALUATION RESULTS

NIST has implemented both of the composite information extraction metrics (Slot error rate and F-measure) for this IE-NE evaluation. F-measure has traditionally been used in the Message Understanding Conferences when calculating entity-tagging performance. The following set of results continues in this tradition and quotes the overall F-measure as the performance metric. Overall F-measure combines the scores from the separate tags ENAMEX, TIMEX, and NUMEX, and weights each according to their frequency.

### 7.1 Baseline IE-NE evaluation

The Baseline IE-NE condition was designed to explore the use of alternative metrics for recognition performance. It also allowed NIST to investigate how differing ASR transcripts (with differing word error rates) would affect the same NE-tagger. It is worth pointing out that different ASR systems produce different errors, hence the success of ROVER. [14] ASR systems that produce transcripts with statistically equivalent word error rates, may produce transcripts that are very different when reviewing the tag-ability of named entities. The nine primary systems had word error rates ranging from 13.5% to 25.7%.

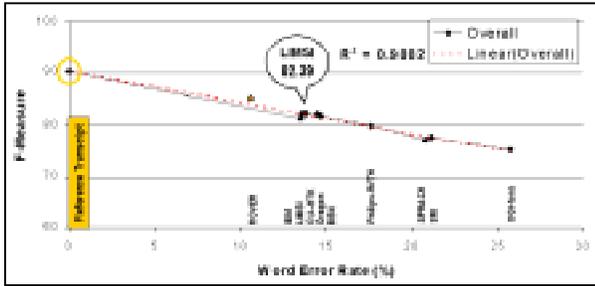


FIGURE 3: Baseline IE-NE Evaluation Results.

Figure 3 shows the results of the Baseline IE-NE evaluation. Running Identifinder on the reference transcript yields an F-measure score of 90.23. The highest F-measure score achieved on the Hub-4 ASR transcripts was 82.39 with the LIMS results. Statistical tests have not yet been implemented on IE-NE evaluation results.

The primary metric for recognition performance is word error rate (WER). Table 1 shows the change in absolute ranking of system performance when the ranking metric is changed from WER to F-measure, and then to slot error rate (SER).

IBM’s relative ranking changes from first by WER to fifth when ordered by either F-measure or Slot Error Rate. The first five systems ranged in performance from 13.5% to 14.7% WER, a difference of only 1.2% WER. The F-measure scores differed by only 0.98 for these same five systems, which leads us to question the significance of this reordering. Analysis of results from the Hub-4 Broadcast News evaluation shows that these 5 systems fall into only two significantly different categories [12] when looking at WER. These same significance tests need to be incorporated into the IE-NE evaluation.

SPRACH and SRI swap seventh and eighth place, respectively.

## 7.2 Quasi IE-NE evaluation

The Quasi IE-NE condition was designed to compare different NE-tagger by having them process identical text streams from the same set of recognizers.

Figure 4 shows the results of the Quasi IE-NE evaluation. Three sites (BBN, SPRACH, and SRI) submitted results for all three baseline recognizers while one (MITRE) submitted results only for baseline recognizer 3.

System	Ranked by WER	Ranked by F-measure	Ranked by SER
IBM	1	5	5
LIMS	2	1	1
CU-HTK	3	2	2
Dragon	4	3	3
BBN	5	4	4
Philips rwth	6	6	6
SPRACH	7	8	8
SRI	8	7	7
OGI fonix	9	9	9

TABLE 1: Shows the change in rank ordering system performance when an alternative IE-NE-based metric is used to determine the ranking. F-measure and SER have the same rank ordering. No Significance tests have yet been performed, table 1 shows the ranking by absolute score.

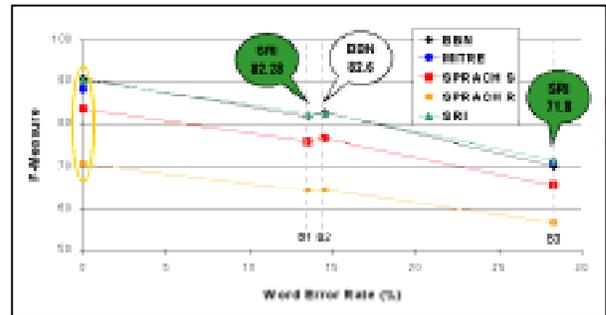


FIGURE 4: Quasi IE-NE Evaluation Results.

SRI achieved the highest F-measure score for baseline recognizer 1 (82.28) and baseline recognizer 3 (71.8). BBN achieved the highest F-measure score for baseline recognizer 2 (82.6) as well as the highest F-measure score when processing the reference transcript (90.56).

## 7.3 Full IE-NE evaluation

The Full IE-NE condition was designed to test the complete information extraction task of one site implementing a speech recognizer and then trying to extract information from that ASR-produced transcript.

Figure 5 shows the results of the Full IE-NE evaluation. Three sites used the same ASR results that their site submitted for the Hub-4 Broadcast News evaluation. MITRE collaborated with SRI International to use SRI’s ASR transcripts for this task.

BBN achieved the highest F-measure score of 82.22 (aided by having the lowest WER transcripts).

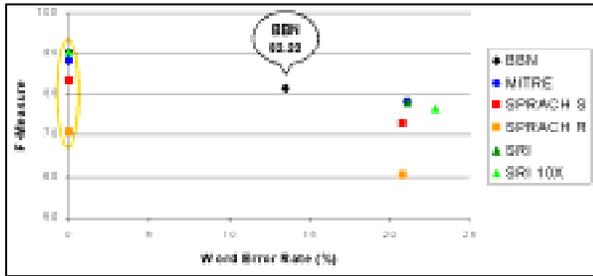


FIGURE 5: Full IE-NE Evaluation Results.

SRI submitted results for two sets of ASR generated transcripts. The first set implemented SRI's NE-tagger on their primary system results, and the second set implemented SRI's NE-tagger on their less than 10X real-time system. Although their less than 10X real-time system had a higher word error rate than their primary system, the performance for entity tagging appears to degrade at the same rate suggesting an equal distribution of errors to both tagged entities and non-tagged entities.

## 8. TAGGED WORD ERROR RATE

It is commonly a goal of system architecture design to have the ability to predict and monitor system performance. One model that was proposed as a predictor of NE performance looked at the word error rate in relation to tagged word tokens and non-tagged word tokens.

Assuming one has the reference transcript available before processing a text stream through a NE-tagger, is it possible to predict the performance?

One method, investigated by NIST, was to look at the WER in the entity fields. It was thought that a tagged word error rate (TWER) might be more strongly correlated with F-measure (or slot error rate) than WER.

To test this hypothesis NIST made use of a meta file (mucscorein) created by **reep**. This file had in it an alignment for each utterance with the named entity tags intact. Using the alignments in mucscorein, NIST generated word error rates for the overall data set, the subset of the data that was not named entities, and the subset of data that was named entities.

It was surprising to find that the WER in the named entity tags (TWER) was higher than that of the WER in the non-named entity tags. Figure 6 shows that tagged words had approximately 20% relative higher word error rates. This goes against accepted folk-lore that automatic speech recognizers have trouble with shorter less informative words (it, the, that, is, in...), but perform stronger on the longer, information-carrying words.

There were ~29,000 word tokens that *were not* marked as being part of a named entity tag. There were only ~2700 words that *were* marked as being part of a named entity tag.

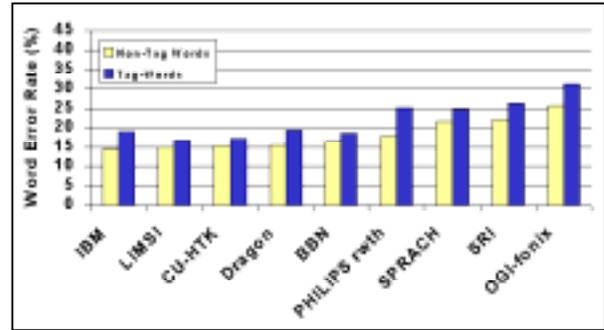


FIGURE 6: Tagged Word Error Rates (TWER). Words that were part of an IE-NE tag had approximately 20% relative higher WER

The strength of the correlation between TWER and the IE-NE metrics turned out to be much weaker than the correlation between WER and the IE-NE metrics. Two explanations may be given for this somewhat unintuitive finding:

1. WER used a much greater sample size than TWER; therefore, WER may be a more stable error metric.
2. TWER was calculated without taking advantage of known NE-tagger insertion errors. Both IE-NE metrics and the WER metric included NE-tagger insertion errors in their calculation. Intuitively, WER, using insertion errors for its calculation, will be more correlated with some other metric that also includes insertion errors in their calculation.

## 9. ROVER TO IMPROVE IE-NE

ROVER has demonstrated significant improvement for ASR results for the past couple of years. The systems used for the Baseline IE-NE evaluation had word error rates that ranged from 13.5% to 25.7%. NIST used ROVER to process these nine systems and produced a result file that had a word error rate of 10.6% [12,14].

Tagging the ROVER results gave an overall F-measure score of 85.30. This point lies above the trendline in figure 3. This suggests that ROVER corrected more entity word errors than non-tagged word errors.

## 10. CONCLUSIONS

The IE-NE evaluation for 1998 was a success in that an objective evaluation of IE-NE technology was developed and implemented. Sites were successful in tagging the information carrying identities as identified in the task definition.

The 1998 Hub-4 Broadcast News Test Set contained plenty of ENAMEX tags, but was light on TIMEX and NUMEX tags as currently defined. With the now proven ability of sites to run ASR systems in less than 10 times real-time without sacrificing word error rate, it seems reasonable to process much larger test sets which would boost the number of TIMEX and NUMEX

tags to create large stable samples.

The linear correlation (and strength thereof) between the IE-NE metrics and WER, suggests that improving recognition will improve entity tagging.

Finally, it was surprising to find that the WER for named entities is higher than that of its compliment data.

## 11. REFERENCES

1. Chinchor, N., Overview of MUC-7, Proc., Message Understanding Conference 7, 1998.
2. Burger, J. D., Palmer, D. D., Hirschman, L., Named Entity Scoring for Speech Input, Proc. 36<sup>th</sup> Annual Meeting of the Association for Computational Linguistics (ACL/COLING '98), August 1998.
3. Chinchor, N., Robinson, P., Brown, E., Hub-4 IE-NE Task Definition Version 4.8,  
[http://www.nist.gov/speech/hub4\\_98/h4\\_iene\\_task\\_def.4.8.ps](http://www.nist.gov/speech/hub4_98/h4_iene_task_def.4.8.ps), August 21, 1998.
5. ALEMBIC Workbench User's Guide,  
<http://www.mitre.org/resources/centers/it/g063/manual2.5/AWB-content.html>
6. Makhoul, J., Kubala, F., Schwartz, R., Performance Measures for Information Extraction
7. R-squared as calculated by Microsoft's linear trend line fit.
8. Evaluation data was annotated by MITRE.
9. The Message Understanding Conference Scoring Software User's Manual,  
<http://online.muc.saic.com/scorer/Manual/manual.html>
11. Douthat, Aaron, SAIC., Scoring Software Ieval0.7,  
[ftp://jaguar.ncsl.nist.gov/csr98/official-IE-98\\_scoring.tar.Z](ftp://jaguar.ncsl.nist.gov/csr98/official-IE-98_scoring.tar.Z)
12. Pallett, D., 1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures, 1999 DARPA Broadcast News Workshop, February 1999.
13. Renals, S., Baseline IE-NE Experiments Using the SPRACH/LASSIE System, 1999 DARPA Broadcast News Workshop proceedings.
14. Fiscus, J., A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER), 1997 IEEE Workshop on Automatic Speech Recognition and Understanding.
15. NIST CTM transcription file format for sclite processing,  
[ftp://jaguar.ncsl.nist.gov/current\\_docs/sctk/doc/infmts.htm#ctm\\_fmt\\_name\\_0](ftp://jaguar.ncsl.nist.gov/current_docs/sctk/doc/infmts.htm#ctm_fmt_name_0).