Robust Named Entity extraction from large spoken archives

Benoît Favre
Thales, MMP Laboratory
Colombes, France
benoit.favre@
fr.thalesgroup.com

Frédéric Béchet
LIA, University of Avignon
Avignon, France
frederic.bechet@
univ-avignon.fr

Pascal Nocéra
LIA, University of Avignon
Avignon, France
pascal.nocera@
univ-avignon.fr

Abstract

Traditional approaches to Information Extraction (IE) from speech input simply consist in applying text based methods to the output of an Automatic Speech Recognition (ASR) system. If it gives satisfaction with low Word Error Rate (WER) transcripts, we believe that a tighter integration of the IE and ASR modules can increase the IE performance in more difficult conditions. More specifically this paper focuses on the robust extraction of Named Entities from speech input where a temporal mismatch between training and test corpora occurs. We describe a Named Entity Recognition (NER) system, developed within the French Rich Broadcast News Transcription program ESTER, which is specifically optimized to process ASR transcripts and can be integrated into the search process of the ASR modules. Finally we show how some *metadata* information can be collected in order to adapt NER and ASR models to new conditions and how they can be used in a task of Named Entity indexation of spoken archives.

1 Introduction

Named Entity Recognition (NER) is a crucial step in many Information Extraction (IE) tasks. It has been a specific task in several evaluation programs such as the Message Understanding Conferences (MUC), the Conferences on Natural Language Learning (CoNLL), the DARPA HUB-5 program or more recently the French ESTER Rich Transcription program on Broadcast News data. Most of these conferences have studied the impact of using transcripts generated by an Automatic Speech Recognition (ASR) system rather than written texts. It appears from these studies that unlike other IE tasks, NER performance is greatly affected by the Word Error Rate (WER) of the transcripts processed. To tackle this problem, different ideas have been proposed: modeling explicitly the ASR errors (Palmer and Ostendorf, 2001) or using the ASR system alternate hypotheses found in word lattices (Saraclar and Sproat, 2004). However performance in NER decreases dramatically when processing high WER transcripts like the ones that are obtained with unmatched conditions between the ASR training model and the data to process. This paper investigates this phenomenon in the framework of the NER task of the French Rich Transcription program of Broadcast News ESTER (Gravier et al., 2004). Several issues are addressed:

- how to jointly optimize the ASR and the NER models ?
- what is the impact in term of ASR and NER performance of a temporal mismatch between the corpora used to train and test the models and how can it be recovered by means of metadata information ?
- Can metadata information be used for indexing large spoken archives ?

After a quick overview of related works in IE from speech input, we present the ESTER evaluation program ; then we introduce a NER system tightly integrated to the ASR process and show how it can successfully index high WER spoken databases thanks to metadata.

2 Information extraction from large spoken archives

The NIST Topic Detection and Tracking (Fiscus and Doddington, 2002) and TREC document retrieval evaluation programs has studied the impact of recognition errors in the overall performance of Information Extraction systems for tasks like story segmentation or topic detection and retrieval. This impact has been shown to be very limited compared to clean text corpora. The main explanation for this phenomenon is the *redundancy effect*: themes, topics are very likely to be represented in texts by many occurrences of salient words characterizing them. Therefore, even if some of these words are missing, numerical Information Extraction methods can use the remaining salient words and discard the noise generated by ASR errors.

However, this phenomenon is not true for tasks related to the extraction of fine grained entities, like Named Entities. Indeed, several studies have shown that F-measure and WER are strongly correlated : 0.7 points of F-measure lost for each additional 1% of WER according to (Miller et al., 2000) on the experiments of 1998 NIST Hub-4 evaluations (Przybocki et al., 1999).

Despite the continuous improvement of ASR techniques, high WER transcriptions are inevitable in difficult conditions like those found in large spoken archives like in the MALACH project (Ramabhadran et al., 2003). Moreover, Named Entities extraction performance is greatly affected by a mismatch between training and testing data. This is due mainly because proper names, which represent most of the Named Entity items, are a very dynamic category of words, strongly related to the period of time representing the documents to process. Therefore this mismatch is inevitable when dealing with archives spreading over a long period of time and containing multiple domain information.

One way of tackling this problem is to gather *metadata* information related to the documents to process. This information can be newspaper corpora related to the same period of time, abstract describing the document content, or simply lists of terms or entities likely to occur. Although such collected data can be used to update the ASR and NER models, the potential gain is rather small unless the metadata corpus gathered fits perfectly the document to process and is of a reasonable size. But another way of exploiting this metadata information is to use it as set of index terms that are going to be explicitly looked for in the processed documents. We present in section 7 an implementation of this idea that uses word lattices as input.

3 The ESTER Named Entity evaluation program

This work has been done within the framework of the French Rich Transcription program of Broadcast News ESTER. ESTER is organized by *l'Association Francophone de la Communication Parlée* (AFCP), *la Délégation Générale pour l'Armement* (DGA) and the *Evaluation language Resources Distribution Agency* (ELDA). The ESTER corpus is made of 100 hours of Broadcast News data (from 6 French speaking radio channels), manually transcribed, and labeled with a tagset of about 30 Named Entity categories folded in 8 main types:

- persons (**pers**): human beings, fiction characters, animals;
- locations (**loc**): geographical, traffic lines, electronic and real addresses, dial numbers;
- organizations (**org**): political, business, non profit;
- geo-socio-political groups (**gsp**): clans, families, nations, administrative regions;
- amounts (**amount**): durations, money, lengths, temperature, age, weight and speed;
- time (**time**): relative and absolute time expressions, hours;
- products (**prod**): art, printings, awards and vehicles;

- facilities (**fac**): buildings, monuments.

This data is divided in 3 sets: a training set (84%), a development set(8%) and a test set (8%). There is a 6 month gap difference between the training corpus and the test corpus while the development corpus matches the training data from a temporal point of view: the training corpus contains Broadcast News spreading from 2002 to December 2003; the development corpus contains news from 2003; the test corpus has been recorded in October 2004. There are also 2 new radio channels in the test corpus which were not in the training data.

For these reasons the development data is called the *matched corpus* as the recording conditions match those of the training corpus and the test data is called the *unmatched corpus*. As a consequence, we can study the effect of unmatched conditions on ASR as well as IE performance and propose solutions for dealing with this problem.

One of the main characteristics of the ESTER corpus is the size of the NE tagset and the high ambiguity rate among the NE categories (eg. administrative regions and geographical locations): 83% of the *matched corpus* entities occur in the training corpus and 40% of them are ambiguous whereas only 61% of the *unmatched corpus* entities occur in the training corpus and 32% of them are ambiguous.

The most commonly used measures for evaluating NE extraction performance are Slot Error Rate (SER) and F-measure. SER is very similar to WER because it takes into account fine grained errors like insertions, deletions and substitutions (entity type and extent). The scoring process is based on the same alignment between reference and hypothesis data than the one obtained for measuring WER and SER is known for being more accurate and penalizing than F-measure. Both measures weights can be adjusted to favor recall or precision and therefore adapted to a specific task.

$$SER = 100 * \frac{\sum_{e \in \mathcal{E}} \alpha_e |e|}{|\textit{Ref slots}|} \qquad F_\beta = \frac{(1+\beta^2)RP}{R+\beta^2 P}$$

$$R = \frac{|\textit{Correct slots}|}{|\textit{Ref slots}|} \qquad P = \frac{|\textit{Correct slots}|}{|\textit{Hyp slots}|}$$

with $e \in \mathcal{E}$ being an error type (insertion, deletion, type, extent, type+extent, multiple) and α_e its weight (resp. $1, 1, .5, .5, .8, 1.5$) ; P is the precision and R the recall; F_1 is used in this paper.

4 Extracting NE from written text vs. ASR output

As previously mentioned in section 2, WER and SER performance are strongly correlated. Besides the intrinsic difficulties of ASR (robustness to noise, speaker variation, lack of coverage of the Language Models used, ...), there is a source of errors which is particularly important in IE from speech input: the Out-Of-Vocabulary (OOV) word phenomenon. Indeed, ASR models are built on huge textual corpus and only contain the most frequent words to limit computation and memory usage. If this is the right approach to WER reduction, it is certainly not valuable to information extraction where unlikely events are considered as important. For instance, many document retrieval models use inverse document frequency (rareness) as a word weighting parameter. So, unlikely proper names are not in reach of the ASR transcription system and hence cannot be spotted by a Named Entity extraction module.

In addition to Out-of-Vocabulary words, two other phenomenons have also a strong impact on NER performance: the insertion of erroneous proper names that automatically trigger the insertion of an entity and spontaneous speech phenomenons. These speech dysfluencies (hesitations, filled pauses, false starts...) reduce the quality of the transcript because they are usually not covered by language models (built from textual data) or artificially introduced. One should remove these from the transcript to improve the quality of the labeling.

In order to deal with ASR errors two approaches have been proposed:

- modeling explicitly the ASR errors, thanks to a development corpus and a set of confidence measures, in order to detect the possible errors of the 1-best word string hypothesis (with the type of errors) before extracting the NEs (Palmer and Ostendorf, 2001);
- exploiting a search space bigger than the 1-best hypothesis alone, either by taking into account an n-best list (Zhai et al., 2004) or the whole word lattice (Saraclar and Sproat, 2004).

The method proposed in this paper is close to this second approach where the whole word lattice output by the ASR system is used in order to increase NER performance from noisy input.

We will present also in the next section a new strategy for adapting NER models to ASR transcripts, based on one of the main characteristics of such transcripts: a closed vocabulary is used by the ASR system. To our knowledge this has never been fully exploited by NER systems. Indeed while the key point of NER systems on written text is their generalization capabilities when processing unknown words, this feature is not relevant for ASR transcripts as the system cannot generate words out of the lexicon (there are no unknown words). Therefore we propose here to fully exploit this constraint (close vocabulary): since the OOV words cannot appear in the ASR transcripts, the NER models can by over-trained on the words belonging to the ASR lexicon. This is going to be developed in the next section.

5 Robust Named Entity extraction

We have developed in this study two NER systems: one is based on the freely available NLP tool *Lingpipe*[1], adapted and trained on the French ESTER corpus and dedicated to process text input. This system is going to be called NER_{text} in the experiment section. The second NER system has been developed for this study and is specifically built for being tightly integrated with the ASR processes. The two main features of this system, called NER_{asr} in the following, are its ability to process word lattices and the fact that the NER models are trained for a specific ASR lexicon. These two systems are going to be presented in the next sections.

5.1 Text-based NER system: NER_{text}

Among all the different methods that have been proposed for NER, one can find rule based models (Cunningham et al., 2002), Maximum Entropy models (Brothwick et al., 1998), Conditionnal Random Fields or probabilistic HMM-based models (Bikel et al., 1999).

Lingpipe implements an HMM-based model. It maximizes the probability of a tag sequence T_i over a word sequence W_i. A context of two preceding words and one preceding tag is used to approximate this probability. Generalization is done through a simple process: words occurring with low frequency are replaced by feature based categories (capitalized, contains digits, . . .). In this approach, there must be one tag per word. Words starting and ending entities are labeled with special tags. Because some features are lacking in ASR transcripts (e.g. capitalization, digits, sentence boundaries, . . .) some word lists for each kind of features are added as presented in (Appelt and Martin, 1999).

5.2 ASR-based NER system: NER_{asr}

Errors occurring in ASR output lead NER systems to overgenerate NE detections. This is due to both erroneous words insertions in the ASR transcripts as well as some abusive generalization made by the NER systems. If these generalization capabilities are very important for processing unknown words in written texts, they can be an handicap in a closed-vocabulary situation like the one observed when processing ASR output. In order to reduce and control the insertion rate of our NER system, we implemented a two level approach: the first level is made of NE grammars coded as Finite State Machine (FSM) transducers and the second level is a statistical HMM-based tagger.

5.2.1 NE transducers

To each NE category is attached a set of regular grammars, extracted from the ESTER training corpus and generalized thanks to the annotation guidelines and web-gathered word lists. Theses grammars are represented by Finite State Machines (FSMs) (thanks to the AT&T GRM/FSM toolkit (Allauzen et al., 2003)). These FSMs are transducers that accept word sequences on the input symbols and output NE labels on the output symbols. They are all grouped together in a single transducer, called T_{gram}, with a filler model that accepts any string of words. Because these FSMs are lexicalized with the words of the ASR lexicon, one can control the generalization capabilities of the grammars thanks to the occurrence contexts of these words in the training corpus. During the NER process, the first step is to compose the FSM representing the NE transducer and the output of the ASR module (either a 1-best word string

[1] Lingpipe: http://alias-i.com/lingpipe/

or a word lattice, both encoded as an FSM called G).

5.2.2 NE tagger

The result of the composition of the NE transducer with the ASR output is an FSM ($G \circ T_{gram}$) containing all the possible parsing made by the NE grammars. In order to find the best analysis a statistical model is used to decide between entity types and entities boundaries. This model is a 2nd order n-gram model (trigram) represented by a weighted FSM (called T_{tagger}) with the same framework as the grammars. The most likely NE label sequence is obtained by finding the best path in the FSM: $G \circ T_{gram} \circ T_{tagger}$. This corresponds to maximize the following probability:

$$P_W = \prod_{i=1}^{n} P(W_i, T_i | W_{i-1}, T_{i-1}, W_{i-2}, T_{i-2})$$

This model is similar to the one implemented in *Lingpipe* but it uses different smoothing methods. Similarly, first and last words of entities are represented by special tags (this helps getting more accurate boundaries) and low frequency words (appearing less than a fixed number of times in the training corpus) are generalized using their Part-Of-Speech tags. The key points of this approach are that it has a better control of the generalization capabilities than a feature based NER system, thanks to the NE grammars; it integrates the closed vocabulary constraint of the ASR systems; and it is not limited to the 1-best word hypothesis but can use the full ASR search space (through word lattices) in order to detect entities. Processing word lattices allows us to output, at the end of the extraction process, an n-best list of NE hypotheses. To each hypothesis are attached two scores:

- the likelihood score given by the ASR model to the best word string supporting this NE hypothesis in the word lattice;
- the probability $P(W_n, T_n, \ldots, W_0, T_0)$ given by the NE tagger to the sequence of NE labels $T_0, \ldots, T_n$ and the sequence of words $W_0, \ldots, W_n$.

From this n-best list we can estimate the *Oracle* performance of the NER system. This measure is the recall measure upper bound than can be obtained by extracting all the possible entities from a word lattice, thanks to the NE transducers, and simulating a perfect strategy that always take the right decision in choosing among all the possible entities.

Decision strategies on such an n-best of NE hypothesis can also involve other levels of information on the document to process like the date or the theme, for example. In the evaluation presented in the next section we compare this Oracle performance measure to the results of the simplest decision strategy which consists in choosing the NE hypothesis with the highest likelihood.

5.3 Evaluation

The evaluation presented in Tables 1 and 2 is performed using the Slot Error Rate and the F-measure on the *matched* and *unmatched* corpora presented in section 3.

corpus	*matched*		*unmatched*	
tagger	SER	F-m	SER	F-m
NER_{text}	21	84	27	79
NER_{asr}	23	84	37	74
WER	0		0	

Figure 1: F-measure and Slot Error Rate measures on the ESTER reference corpora (*matched* and *unmatched*) for both NER systems

corpus	*matched*		*unmatched*		
tagger	SER	F-m	SER	F-m	Oracle
NER_{text}	42	72	55	63	61.9
NER_{asr}	41	73	54	63	76.9
WER	21.2		26.4		

Figure 2: F-measure, Slot Error Rate and Oracle recall measures on the ASR output of the *matched* and *unmatched* corpora for both NER systems

Figure 1 presents SER and F-measure on the two test sets (*matched* and *unmatched*) for the text oriented (NER_{text}) and the speech oriented (NER_{asr}) NER systems, on clean text (manual transcripts). Figure 2 shows the results obtained on the ASR transcripts.

As expected on manually transcribed data, NER_{text} obtains better results than NER_{asr} (which has poorer generalization capabilities). On the ASR outputs the results obtained by both systems are comparable however NER_{asr} has the advantage of processing word lattices, leading to an interesting Oracle performance. We are studying now more elaborate decision strategies in order to take fully advantage of this feature.

The decrease in F-measure observed between the reference and the ASR transcripts is similar to the one obtained in other studies (Miller et al., 2000). One observation that can be made on these results is the impact of the time mismatch between the training and the test corpora. A 6 month difference in the *unmatched* corpus leads to a very big drop in both SER and F-measure. This can be explained by the fact that NEs are very time-dependent. We are going now to present some methods developed to tackle this problem.

6 Updating Language and NE models with metadata information

The only mismatch between the training and the *unmatched* corpus of our experiments is a 6 months temporal mismatch, therefore we collected a corpus of newsletters made on a daily basis by the French newspaper *Le Monde* corresponding to these 6 months. These newsletters contain an abstract of the news of each day. We make the following two hypotheses:

- firstly these newsletters are related to the same time period as the *unmatched* corpus, therefore integrating them into the ASR models (lexicon+Language Model) should help reducing the OOV word effect;
- secondly because they represent an abstract of the news of each day, the Named Entities occurring in a particular newsletter should contain all the major events of the corresponding day and therefore constitutes a useful list of terms that can be used for indexing a Broadcast News document related to the same period of time.

6.1 NE distribution analysis

This newsletter corpus contains 1M words and after being tagged by the NER_{text} system, 140k entities were extracted. To check the relevance of this corpus for adapting the models to the *unmatched* test corpus, we studied the distribution of the words and the entities for each day, from January to December 2004. The *unmatched* test corpus is made of Broadcast News ranging from October 10th to October 16th 2004. The following observations were made: 72% of the NEs and 60% of the words contained in them occur only one day in this corpus; the intersection of the NEs occurring in both the newsletter of a particular day and the entities belonging to the *unmatched* test corpus shows a peak, illustrated by figure 3, for the days of the test corpus; at this peak, 25% of the NEs are used the same day in the two corpora.

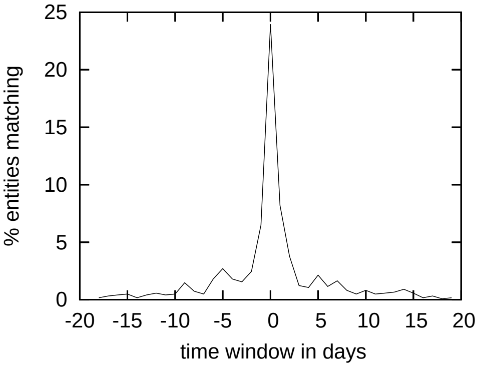

Figure 3: Percentage of entities of the unmatched corpus occuring at least n days earlier or later in the newsletter corpus (at a window of 0 days, entities appear on the same day in both corpus).

The first observation matches those presented in (Whittaker, 2001) and validates our approach which consists in carefully adapting the ASR and NER models with data corresponding to the exact period of time as the one of the documents to process: by taking into account a larger period of time for the adaptation corpus, the necessity of restraining the models to the most frequent entities would lead to discard low frequency terms that can be crucial for characterizing the news of a given day.

If the second observation clearly highlights the correlation between the NE distribution in both corpora, it also points out that only 25% of the entities of the *unmatched* corpus occur in the newsletters corresponding to the same days. Therefore the potential improvement in the overall NER performance is clearly limited. This will be confirmed in the next section, however one can think that if these entities are shared, for a given day, by both corpora, it is because they represent the key topics of this day and therefore they can be considered as very relevant indexing terms for applications like document retrieval. This last point is developed in section 7.

6.2 Model adaptation

Several studies (Whittaker, 2001; Federico and Bertoldi, 2001; Chen et al., 2004) propose adaptation methods of a general language model to the possibly small corpora extracted from these kinds of metadata information (an overview of these methods can be found in (Bellegarda, 2004)). Depending on the adaptation method and the kind of metadata information used, some gains in performance have been reported. But it appears that the choice of the metadata and the size of the adaptation corpus collected are critical in this adaptation process: if the adaptation corpus is not exactly related to the topics of the document to process, no real gains are obtained (e.g. (Chen et al., 2004) reports that the best gains were obtained with a story-based adaptation method).

From all these previous works, our system implements the following adaptation process:

- the text corpus corresponding to the newsletters is added to the ASR language model by means of a linear interpolation;
- proper names occuring twice or more in the newsletter corpus are added to the ASR lexicon;
- for the same days as those of the *unmatched* corpus, this cutoff is suppressed and all the proper names are added;
- the Named Entity wordlists and grammars are also enriched with these proper names and entities extracted from the collected corpus.

1K new proper names were added to the 65K word ASR lexicon. The general OOV reduction obtained was 0.14% leading to an absolute WER reduction of 0.3%. Similarly the SER decreased of about 0.3% thanks to this adaptation and the Oracle recall measure in the word lattices was improved by an absolute 3%. These improvements are not significant enough to justify the use of this kind of metadata information for improving the general performance of both ASR and NER processes. However, if we focus now on the entities occurring in the newsletters corresponding to the exact days of the *unmatched* corpus, the improvement is much more significant, as presented in the next section.

7 Named Entity Indexation

As previously mentioned, 25% of the *unmatched* corpus entities occur in the newsletters corresponding to the same day as those of the *unmatched* test. In order to measure the improvement obtained with our adaptation technique on these particular entities, we did the following experiment:

- a set of 352 entities was selected from the newsletters related to same period of time as the test, these entities represent the indexing terms that are going to be looked for in the word lattices of the *unmatched* corpus;
- the NER_{asr} system was then applied to these word lattices with two conditions: the word lattices and the NER models before adaptation and those obtained after adaptation with the newsletter corpus;
- precision, recall, F-measure and Oracle error rate were estimated for both conditions.

Condition	Prec.	Recall	F-m	Oracle
no adaptation	87.0	75.7	80.9	83.6
with adaptation	87.5	83.9	85.7	92

Figure 4: Extraction results on the selected NEs on the *unmatched* corpus with and without adaptation of the ASR and NER models on the newsletter corpus

As we can see in table 4, the adaptation process increases very significantly the recall measure of the NE extraction. This is particularly relevant in some IE tasks like the document retrieval task.

8 Conclusion

We have presented in this paper a robust Named Entity Recognition system dedicated to process ASR transcripts. The FSM-based approach allows us to control the generalization capabilities of the system while the statistical tagger provides good labeling decisions. The main feature of this system is its ability to extract n-best lists of NE hypothesis from word lattices leaving the decision strategy choosing to either emphasize the recall or the precision of the extraction, according to the task targeted. A comparison between this approach and a standard approach based on the NLP tools *Lingpipe* validates our hypotheses. This integration of the ASR and the NER processes is particularly important in difficult conditions like those that can be found in large spoken archives where the training corpus does not match all the documents to process. A study of the use of metadata information in order to adapt the ASR and NER models to a specific situation showed that if the overall improvement is small, some salient information related to the metadata added can be better extracted by means of this adaptation.

References

Cyril Allauzen, Mehryar Mohri, and Brian Roark. 2003. Generalized algorithms for constructing statistical language models. In *ACL'03, Sapporo, Japan*.

D. Appelt and D. Martin. 1999. Named entity extraction from speech: Approach and results using the TextPro system. In *Proceedings Darpa Broadcast News Workshop*.

Jerome R. Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42 Issue 1:93–108.

Daniel M. Bikel, Richard L. Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. volume 24, pages 211–231.

Andrew Brothwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Exploiting diverse knowledge sources via maximum entropy in named entity recognition.

Langzhou Chen, Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. 2004. Dynamic language modeling for broadcast news. In *In International Conference on Speech and Language Processing*, pages 1281–1284.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

M. Federico and N. Bertoldi. 2001. Broadcast news LM adaptation using contemporary texts. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech)*, pages 239–242, Aalborg, Denmark.

Jonathan G. Fiscus and George R. Doddington. 2002. Topic detection and tracking evaluation overview. *Topic detection and tracking: event-based information organization*, pages 17–31.

G. Gravier, J.F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri. 2004. ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. In *Proc. Journées d'Etude sur la Parole (JEP)*.

David Miller, Sean Boisen, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. 2000. Named entity extraction from noisy input: Speech and OCR. In *Proceedings of ANLP-NAACL 2000*, pages 316–324.

D. D. Palmer and M. Ostendorf. 2001. Improving information extraction by modeling errors in speech recognizer output. In *Proceedings of the First International Conference on Human Language Technology Research*.

M. A. Przybocki, J. G. Fiscus, J. S. Garofolo, and D. S. Pallett. 1999. 1998 Hub-4 Information Extraction Evaluation. In *Proceedings Of The DARPA Broadcast News Workshop*, pages 13–18. Morgan Kaufmann Publishers.

Bhuvana Ramabhadran, Jing Huang, and Michael Picheny. 2003. Towards automatic transcription of large spoken archives - english ASR for the MALACH project. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 216–219.

Murat Saraclar and Richard Sproat. 2004. Lattice-based search for spoken utterance retrieval. In *HLT-NAACL 2004: Main Proceedings*, pages 129–136, Boston, Massachusetts, USA. Association for Computational Linguistics.

E. W. D. Whittaker. 2001. Temporal adaptation of language models. In *Adaptation Methods for Speech Recognition, ISCA Tutorial and Research Workshop (ITRW)*, August. LM Adaptation for information retrieval of spoken news/radio programs (i.e. SpeechBot).

Lufeng Zhai, Pascale Fung, Richard Schwartz, Marine Carpuat, and Dekai Wu. 2004. Using n-best lists for named entity recognition from chinese speech. In *HLT-NAACL 2004: Short Papers*, pages 37–40, Boston, Massachusetts, USA. Association for Computational Linguistics.