Improving Out-of-Vocabulary Name Resolution

David D. Palmer
dpalmer@virage.com
Advanced Technology Group, Virage Inc.
Woburn, MA 01801

Mari Ostendorf
mo@ee.washington.edu
Dept of EE, University of Washington
Seattle WA, 98195-2500

UW
Electrical
Engineering

UWEE Technical Report
Number UWEETR-2003-0005
March 2003

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

Improving Out-of-Vocabulary Name Resolution

David D. Palmer
dpalmer@virage.com
Advanced Technology Group, Virage Inc.
Woburn, MA 01801

Mari Ostendorf
mo@ee.washington.edu
Dept of EE, University of Washington
Seattle WA, 98195-2500

University of Washington, Dept. of EE, UWEETR-2003-0005

March 2003

Abstract

This paper presents algorithms for generating targeted name lists for candidate out-of-vocabulary (OOV) words for applications in language processing, particularly speech recognition. Focusing on names, which are shown to be the dominant class of OOVs in news broadcasts, the approach involves offline generation of a large name list and online pruning based on a phonetic distance. The resulting list can be used in a rescoring pass in automatic speech recognition. We also show that a simple variation of the approach can be used to generate alternate name spellings which may be useful for query expansion in information retrieval. By using a wide variety of sources, including automatic name phrase tagging of temporally relevant news text, OOV coverage can be improved by nearly a factor of two with only a 10% increase in the word list size. For one source, coverage increased from 13% to 94%. Phonetic pruning can be used to reduce the list size by an order of magnitude with only a small loss in coverage.

1 Introduction

In many language processing systems, the vocabulary – the specific words that a system is capable of processing – is predetermined and finite. For example, an automatic speech recognition (ASR) system has a vocabulary of words that it can produce, and any spoken word that is out-of-vocabulary (OOV) is mapped to the closest in-vocabulary (IV) word or words. Similarly, an information retrieval (IR) system can only index the words contained in the documents from which it is trained; the IR system will be unable to return relevant documents for queries containing OOV words.

This out-of-vocabulary problem in language processing has led to the need for techniques that robustly process data containing OOV words. In the case of information retrieval, OOV words often result from misspellings or spelling variants of known words, and the OOV problem can be addressed by term normalization, mapping OOV words to similar words contained in the index; however, this mapping between similar words is not always straightforward. In automatic speech recognition, the OOV words are much harder to identify, since most ASR systems always produce a hypothesized word from a fixed vocabulary. In either case, the OOV error correction and term normalization problems are similar: an orthographically- or phonetically-similar variant must be identified for each OOV word.

As explored in further detail later, the OOV problem is quite important, particularly for names. Hetherington [18] found that, in speech recognition, each OOV token contributes on average 1.5 errors, in part because longer OOVs (such as names and places) are often recognized with multiple short, high frequency words. In addition, he found that 20% of the errors resulting from OOV words correspond to in-vocabulary words being misrecognized due to their proximity to an unknown word. Even though the use of a very large vocabulary in a recognition system can bring the OOV rate below 1% in broadcast news data, the number of sentences having at least one OOV word is still high, roughly 10-15% with a 57k dictionary for the news data used in this work. In highly inflected languages or those with a high rate of word compounding, the OOV rates tend to be even higher [14]. In English, the OOV rate is especially high for the words with the most information content: in the news data analyzed here, 45.2% of OOV word tokens are in person phrases, although person words make up only 3% of all tokens.

In speech recognition, the simplest method for addressing the OOV word problem is to increase the vocabulary size of the system. For example, Watclar *et al.* [38] developed a method for supplementing a broadcast news ASR system vocabulary using words from recent text-based news sources. However, systems with larger vocabularies require more memory and run slower than those with smaller vocabularies. Since practical ASR systems cannot have unlimited memory and computational requirements, they naturally cannot have unlimited vocabulary sizes. In addition to increased computational cost, adding words to a vocabulary increases the potential confusability with other vocabulary words. Rosenfeld [33] reports that a vocabulary size of around 64k is nearly optimal for processing read North American Business news text, and that increasing the vocabulary size beyond this yields negligible recognition improvement at best. Of course, the optimal vocabulary size is also domain dependent: a 64k word vocabulary may not be necessary for travel dialog but may be inadequate for directory assistance. Rosenfeld's analysis shows that increasing the system vocabulary size can help recognition rates for many common words while hurting rates for less common words. Yet the less common words, such as new names introduced as a result of national and international events, usually contain more semantic information about the utterance, and these errors are much more costly for language understanding applications.

An important problem in vocabulary design, which can be independent of size limitations, is identifying and ranking the most important vocabulary items. Since new words are constantly being introduced into common usage, the training data used to build an ASR system or a set of document indexing terms will generally not cover all vocabulary of interest in future use of the system. It is necessary to identify alternate lexical resources that contain words likely to be of interest. Of course, the appropriate strategy will depend somewhat on the domain. In our work, the focus is on broadcast news and the name problem in particular, so the sources are motivated by that application. However, in addition to judicious choice of sources, we show that text filtering can greatly help improve vocabulary coverage, rather than relying simply on word frequency.

In this paper, we present the results of our work in correcting errors resulting from out-of-vocabulary words, focusing on person names. Our approach to name list generation is to use extensive lexical resources combined with a text-based information extraction system to maximize vocabulary coverage. We introduce an ASR post-processing approach that allows us to treat error correction as a type of spelling correction task, and we show how a phonetic distance calculation can be used to further prune the candidate name list by ranking the items in a vocabulary list according to their similarity to the hypothesized word. We also illustrate in a pilot experiment how the same modules can be used for name normalization in text processing, which has implications for information retrieval and extraction applications. Analyses and experiments are based on speech and text data collected for topic, detection and tracking (TDT) research. The TDT corpus [17] includes news broadcasts from 1997-1998, transcribed by Dragon for the TDT evaluation, as well as large collections of text data from the New York Times (NYT) and Associated Press Newswire Service (APW) collected at roughly the same time. For the work described in this paper, we used 114 broadcasts from the TDT corpus, consisting of about 730,000 words and 35,000 name phrases.

ASR Output	*the shirts show our straws year behind it with a sledgehammer and a racist caption*
ASR With Name Labels	*the shirts show* [location] [person] *with a sledgehammer and a racist caption*
Reference Transcription	The T-shirts showed Austria's Jörg Haider with a sledgehammer and a racist caption

Figure 1: Example showing the information content of names: ASR output (30% WER) for a sentence, the same ASR sentence with regions containing name phrases labeled, and the correct transcription.

In Section 2 we provide a detailed analysis of the out-of-vocabulary word problem and motivate our focus on OOV names. In order to put our work in context, in Section 3 we review previous work in OOV detection and error correction, as well as spelling correction. In Section 4, we describe our general framework for OOV processing in the context of the ASR and name normalization problems. The two main components of the approach – list generation and pruning via phonetic distance – are described in Sections 5 and 6, respectively. Section 6 also contains empirical results of the application of our approach to the problems of list pruning, ASR error correction, and name normalization. We conclude with a discussion of remaining problems in Section 7.

2 The Problem of OOVs and Names

Name phrases, specifically names of persons, locations, and organizations, contain a great deal of information about the sentence in which they appear, and they occur quite frequently in many sources. In the TDT broadcast news corpus, 9.4% of the words are contained in name phrases, and 45.1% of the utterances contain at least one name phrase. Names are also very frequently the source of word errors in the ASR output; in the broadcast news data, the word error rate for words within name phrases is 38.6% while the word error rate for non-name words is 29.4%. This combination of high density and high error rate results in a great deal of information being lost due to name word errors, specifically out-of-vocabulary words.

Figure 1 shows an example from a November 2000 news broadcast and illustrates the importance of names in a news story. Three versions of a sentence from the news are shown. In order to demonstrate the importance of names, the order is reversed such that each version provides more information than the previous. The first version shows the actual raw ASR output for a sentence; due to the word errors "*our straws year behind it*," it is extremely difficult to extract any information from the sentence. However, the second version in Figure 1 shows that if we know that "*our straws*" is a location phrase and that "*year behind it*" is a person phrase (albeit incorrectly transcribed), we gain a great deal of information. Finally, in the third version, the "clean" reference transcription, we have the complete information, with the names correctly transcribed. A key motivation for the work here is the fact that we can extract information from the ASR output transcription, even when it contains output word errors, by labeling regions with high semantic content, such as names. Furthermore, correcting the ASR errors in the names allows us to extract even more information.

Proper names are an important focus of error correction work for the simple reason that there is a strong correlation between names and OOV words. While the overall out-of-vocabulary rate is typically very low (less than 1%) for most large-vocabulary (48k-64k) recognition systems, the OOV rate is significantly higher for words in name phrases. Table 1 shows the OOV rates in the TDT broadcast news data for non-name words and for name words, for a system with a 57k word vocabulary. The results are shown for both **word tokens**, in which all occurrences of a word are counted, and **word types**, in which only unique words

Word category	OOV Rate (%)	
	Word Tokens	Word Types
Non-Name	0.9	6.8
Name Phrases	6.0	27.1
Person names	12.8	34.2
Location names	3.2	17.7
Organization names	1.4	9.0

Table 1: *OOV rates for non-names vs. names, and for specific categories of names.*

are counted.[1] The first two rows show that only 0.9% of non-name word tokens and 6.8% of non-name word types in the data are OOV, while 6.0% of word tokens and 27.1% of word types within name phrases are OOV. The remaining rows show the OOV rates for the name phrase categories. The rows for person, location, and organization names demonstrate the range of OOV rates for the name phrase categories and show that the OOV rates are higher for all name categories than for non-name words. In fact, although name words represent 9.4% of the word tokens and 13.5% of the word types in the TDT data, they account for 57.6% of the out-of-vocabulary word tokens and 43.3% of the OOV types. However, the statistics are more striking for person names: 45.2% of all OOV word tokens and 32.1% of types in the TDT data are in person phrases, although person words make up only 3% of all tokens.

It is extremely difficult to anticipate new names of global importance that will enter the news. For example, in the front page headlines of the Boston Globe newspaper for the first three weeks of January 2001, there were several names of world leaders – Haider (Austria), Kostunica (Serbia), Putin (Russia), Macapagal (The Philippines) – that did not occur a single time in the broadcast news portion of the TDT corpus. Kostunica also did not appear in the NYT/APW portion, and the other three appeared in the print sources only a few times each (Putin one time, Macapagal seven times, Haider eighteen times). There are always new names entering the news, and we will never be able to explicitly model all of them in an ASR vocabulary; however, we lose a great deal of information when the names of world leaders (in particular) are OOV.

Many instances of out-of-vocabulary words in applications such as speech recognition, information extraction, and information retrieval are actually very similar to in-vocabulary words, but differ slightly due to spelling or morphological variations (both intentional and accidental). For example, the word “Coburns” in our ASR data is OOV although “Coburn” is in-vocabulary; similarly, “Lewinsky” is OOV although “Lewinski” is IV. In the case of information retrieval, spelling errors and spelling variants in search terms (names, in particular) are quite common; for example, there are over 50 valid variants of the name of the leader of Libya (e.g., Gadhafi, Qadafi, Khadafi). In the case of information extraction, it can be difficult to automatically identify and link salient information when spelling variants or misspellings refer to the same person. In all cases, it is important to identify these OOV variants and to normalize all instances that refer to the same entity. Contextual information may be needed to distinguish variants of the name of a single entity, like Gadhafi, from similar names that refer to different entities, like (George) Lewinski and (Monica) Lewinsky, but identifying the candidates is a necessary first step in either case.

3 Related Work

The research described in this paper builds on and impacts work in a wide range of research areas. In this section we provide an overview of previous work, with a particular focus on the two main categories, OOV

[1] For example, the sentence “The bat hit the ball” has five word tokens and four word types.

detection/correction and spelling correction.

3.1 OOV Detection and Correction in ASR

There has been some attention in recent years to the general problem of finding OOV words during speech recognition. Most of the work has focused on detecting out-of-vocabulary words by developing garbage models: acoustic and language models of a generic unknown word, ideally dissimilar to all words in an ASR system's existing vocabulary. Asadi *et al.* [2] present one of the first attempts to model OOV words in a continuous speech recognition system by detecting new words and adding them to the ASR vocabulary. They develop and compare several HMM acoustic models of new words, with each model explicitly representing the sequence of phonemes in the new words within the speech recognizer. They also use a class grammar to restrict the appearance of new words to open classes, such as ship names and port names. Extensions to this approach have included both changes to the structure of the acoustic "garbage" model and more sophisticated language models; see, e.g., [36, 5, 13, 4, 34].

In addition to work in detecting OOV words, some previous work has also attempted to transcribe or correct ASR errors. Chung [9] presents a three-stage method for detecting out-of-vocabulary city names in a weather information system and also describes a technique for phonetically and orthographically transcribing the unknown city names by explicitly modeling both graphemes and phonemes in the pronunciation modeling. The approach uses a 2000 word vocabulary trained on 56k sentences, of which 2000 contained unknown city names (annotated as such); it was tested on 425 utterances, each containing exactly one OOV city name and was reported to identify 81.4% of the OOV city names.

Geutner *et al.* [15] describe a multi-pass ASR decoding approach targeted at reducing the out-of-vocabulary rates for heavily inflected language, such as Serbo-Croatian, German, and Turkish. Their work attempts to dynamically expand the effective vocabulary size by adapting the recognition dictionary to each utterance. In the first recognition pass, an utterance-specific vocabulary list is constructed from the word lattice. They then use a technique they call "Hypothesis Driven Lexical Adaptation" to expand the vocabulary list by adding all words in a full dictionary that are sufficiently similar to those in the utterance list. Geutner *et al.* report that the lexical adaptation methods result in a significant decrease of up to 55% in OOV rates for the inflected languages.

An ASR post-processing approach that focused on in-vocabulary errors was that of [31], who developed a post-processor called SpeechPP for correcting word-level ASR errors in speech recognition output for applications where the ASR system is used in a new application, as when an ASR system trained for the air travel domain is applied to the train travel domain. Taking advantage of the fact that a recognizer makes consistent errors in such a domain change, they model the error process as a noisy channel combined with a fertility model to account for the fact that not all recognizer errors are one-to-one word errors. They then use a Viterbi search to produce a word sequence that is more likely than the actual ASR output, resulting in a word error rate reduction as high as 24.0% in the train travel domain.

In comparing our work to the other work with OOV words in ASR systems, we emphasize several important aspects. Our approach is an ASR post-processing model, in which we do not require an acoustic model. In contrast, much of the previous work has focused on explicitly modeling errors within the ASR process itself. Our work is thus complementary to previous approaches, yet has the advantage of not requiring that the audio signal be archived. In addition, our models make no assumptions about the position or frequency of errors in the training or test data. For spoken document information retrieval applications, a different post-processing approach has been proposed that maps queries to possible IV word combinations [24], which is similar to our approach in not requiring re-recognition but it does no correction and hence is less useful for information extraction.

3.2 Spelling Correction

The correction of spelling errors in text documents has been a topic of research in the natural language processing and machine learning communities for decades. The primary goal of spelling correction is to identify and replace words in a text document that have been either mistyped (e.g., *teh* for *the*) or confused with a similar word (e.g., *piece* for *peace*). Here we briefly summarize key methods that relate to our approach; see Chapter 5 of [20] for a more thorough discussion of spelling correction techniques.

One of the most successful approaches to general-purpose spelling correction has been the noisy channel error model. In this approach, misspellings are assumed to have been generated by transmission through a channel that corrupts the original word. Noisy channel spelling correction consists of models of the types of word that might be sent through the channel and of the characteristics of the channel that generates the spelling errors. One common channel error model uses a variation of the Levenshtein distance [23] – the minimum number of character insertions, deletions, and substitutions – required to convert the candidate misspelling into a certain known word. For example, converting the string *fysics* into *physics* would require a minimum of two edits, a substitution and an insertion. In developing the noisy channel model for spelling correction, individual character edits have been weighted equally [25], and with weights of different magnitudes [10, 21]. In the case of variable weights, weights can be assigned heuristically or empirically, from a training corpus of known misspellings.

Ristad and Yianilos [32] proposed a further improvement in edit weight estimation. They developed a stochastic transduction model, where the individual string edit operations are hidden events. They then used the expectation-maximization algorithm [11] to derive edit weights based on all possible edit sequences for a given pair of strings. They showed that this weight estimation algorithm produced a significant improvement in spelling correction performance, compared with the Levenshtein distance both with uniform weights and with weights derived from only the most likely edit sequence between string pairs.

In all previous work, string edits for spelling correction had been limited to simple character-to-character substitutions and single character insertions/deletions. Brill and Moore [6] developed a further improvement on the noisy channel spelling correction model by allowing more general edits, such as the substitution of a sequence of characters for another sequence. With this model, converting the string *fysics* into *physics* would require only one edit, a substitution of the characters *ph* for *f*. The set of possible edits and the edit weights are both empirically learned from a training corpus of known misspellings. Brill and Moore report a 52% performance improvement using this general model, compared to the weighted Levenshtein distance used by [21].

Although we can (and do) build on these results and use similar techniques to correct sequences of phonetic symbols, we should emphasize that ASR error correction or normalization is a significantly more difficult problem than spelling correction. ASR errors are much more difficult to identify as such: many misspellings can be detected simply by the fact that the word is not contained in an extensive dictionary, while for most systems the ASR output words are **always** in the lexicon. In spelling correction, there is quite often a one-to-one mapping between the hypothesized (misspelled) word and the corrected word; however, a single spoken word often results in multiple ASR output errors. Furthermore, the types of errors that humans most commonly make in misspellings, transpositions (*teh* vs. *the*) and single-letter substitutions (*reluctent* vs. *reluctant*), are easier to model than the types of errors that ASR systems make in mistranscriptions. For example, a human may write “Britany Speers” or “Britny Speres,” which are still easily identified as misspellings of “Britney Spears.” An ASR system may instead produce such sequences as “Britain ease peers” or “bread knees beers,” which are significantly more difficult to identify.

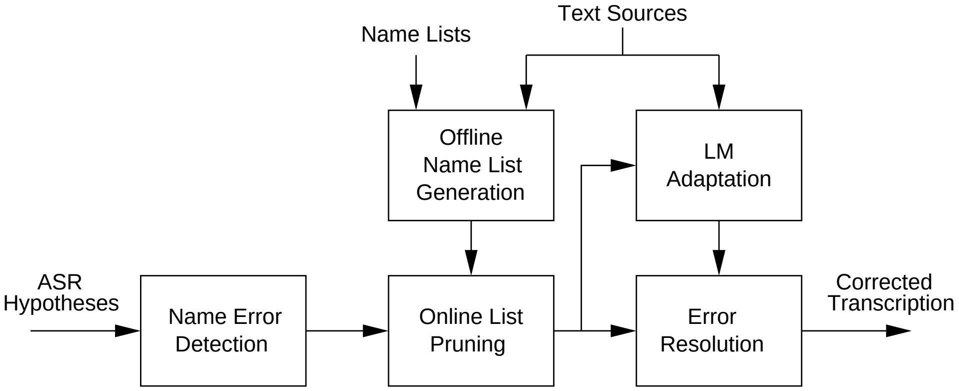

Figure 2: Block diagram illustrating name resolution modules.

4 General Modules and Application Contexts

There are several key components to our error correction approach, as illustrated in Figure 2 for correction of OOV names in ASR. Offline processing involves collecting text sources (news and name lists) and filtering these to build a new list of names and their associated pronunciation for use in online processing. The first step in online correction is to identify candidate tokens for error correction or normalization, referred to in the figure as “OOV detection.” Then, the phone sequence associated with the target word is compared to the pronunciations of words in the full list to generate a pruned list of candidates based on a phonetic distance. The final stage selects from among these candidates, optionally using another pass of recognition. The general framework is somewhat similar to that of [15] except that our focus is on names rather than morphological word differences. Hence, detection is different, and it is possible to get improved error rates using only the phonetic distance and language cues rather than a full acoustic rescoring. Each of these steps is described briefly below, and the two focus problems – list generation and phonetic pruning – are expanded on in the two subsequent sections.

In name normalization, the same types of modules would be useful, but the detection and rescoring mechanisms would obviously differ. Since the focus of this work is on the list generation and phonetic pruning modules, we also include a pilot study (in Section 6.4) showing how the phonetic distance can be used to identify alternative spellings.

Offline Word List Generation. The process of word list generation involves identifying good lexical resources, which may include name lists as well as text sources, then filtering the text sources, and ranking words based on frequency statistics obtained from the text sources. For broadcast news transcription applications, where names are the most important category of OOVs, our approach to text filtering involves automatic name detection, as described further in Section 5. For this and other applications, one could also imagine filtering (or weighting) text based on measures of document relevance, as in [19]. Once the list is generated, phoneme-based pronunciation strings are produced for each word, for use in the online scoring stage. This can be done automatically using a grapheme-to-phoneme converter.

OOV Candidate Detection. As described in Section 3.1, there are many alternatives for OOV detection, based on acoustic cues and/or ASR error patterns. Not surprisingly, performance tends to be better when more information sources are used, and because of our focus on names it is relevant to consider the surrounding language context. Hence the work here uses an approach described in [28] that involves the integration of word confidences into a probabilistic model that can jointly identify names and errors. We construct a simple lattice from the hypothesized word sequence, with “error” arcs in parallel with each word and probabilities associated with each arc corresponding to word confidence estimates [16, 27]. We then use Viterbi decoding to find the path through the lattice that maximizes the likelihood of the joint word/error and name-state sequence according to a statistical information extraction (IE) model trained to recognize names [26]. Although we obtained good information extraction performance using the system, the error detection performance is rather poor. The error detection performance for the test data, using a confidence thresholding technique, is 66.1% recall and 48.8% precision. (Note that, for OOV correction, recall is more important than precision, since the correction step can involve leaving the hypothesized word unchanged.) While these results seem much lower than other reported results, such as those described in Section 3, it is important to note that our work is carried out using a significantly larger vocabulary size than most previous work in OOV detection, and that some of the reports are based on a “biased” test set of utterances known to contain OOV words.

Online List Pruning. Once a word or word sequence in the recognition output is identified as a candidate OOV, the phone sequence for that word is obtained from the recognition dictionary. This pronunciation sequence can be compared to the pronunciations for each of the words in the extended word list, and a distance computed using a string matching procedure and a set of phone substitution, insertion and deletion costs. The alternative words are then ranked according to distance, and optionally word frequency in the case of ties. The phonetic distance and the impact of pruning on list size is described further in Section 6.

OOV Resolution. There are several alternatives for selecting the final word choice from the pruned set of candidates, including using the phonetic distance alone or in combination with a language model score, or via an additional pass of recognition. In our work, where a complete recognition system was not available, the selection is based only on the phonetic distance and the reduction in error rate is small. While the gains would likely be larger by rerunning recognition with the expanded vocabulary, there are many applications (such as in information retrieval from a large corpus), where subsequent recognition runs are impractical. Whether using the phonetic distance alone or rerecognition, it is likely that using an adapted language model based on temporally or topically relevant text containing the target words will be important to achieving high accuracy, particularly for resolving spelling alternatives (e.g. “Lewinsky” vs. “Lewinski”). This emphasizes a significant advantage of the post-processing approach to OOV detection: we have valuable hindsight about the context in which candidate OOVs appeared.

5 Vocabulary Coverage and Name List Generation

Just as an ASR system vocabulary is chosen to provide maximal coverage of words it might encounter, for OOV error correction we need to identify lexical resources that provide the maximal coverage of OOV names in the data. In this section, we present the results of our analyses in determining some factors that contribute to optimal vocabulary coverage. Since we also want to minimize the size of the candidate list to allow efficient search, we also discuss the tradeoff between list size and coverage.

5.1 Analysis of OOV Names in Broadcast News

In Section 2, we showed that 45.2% of all OOV words in broadcast news data are person names, compared to just 7.4% for location names and 4.6% for organization names. In addition, while there is a virtually endless supply of new person names entering the news, the number of new locations is fairly small, and new organizations are very often composed of person and location names. For these reasons, we focus our analysis on coverage of person names, and our error correction results will be limited to person names.

As an initial data point in vocabulary coverage, we can use large name lists, such as the list of over 90,000 surnames and first names made available to the public in 1995 by the U.S. Census Bureau [37], based on the 1990 Census. The Census data provides an extensive list of the most frequent person names in the United States, ranked in order of frequency. We examined the vocabulary coverage provided by the Census name list for both **word tokens**, in which all occurrences of an OOV word in the news are counted, and **word types**, in which only a single occurrence of each OOV word is counted. The Census list provides surprisingly poor coverage, containing only 31.4% of the OOV tokens and 29.1% of the OOV types. There are several possible reasons for this poor coverage. To maintain the anonymity of the surveys collected by the Census Bureau, they truncated the list made available to the public so that it does not include rare surnames that might allow a particular Census form to be immediately linked to an individual or family.[2] The truncated list contains 88,799 names that represent 90% of all individuals surveyed, but there are hundreds of thousands of surnames in the tail of the distribution that are not contained in the Census lists. Another reason for the poor coverage is the fact that many names in the news are foreign names that might not be represented in the Census data; Appelt and Martin [1] report using a (proprietary) list of names of many nationalities that provides stronger coverage of names in the news than the Census list.

In order to identify additional lexical resources that might help increase the vocabulary coverage, we examined the OOV person words from the TDT training data. Our analysis indicated there were five primary sources of OOV person names that require a variety of name sources to increase overall coverage.

"New" Names of Global Importance. This is perhaps the most important category of OOV names in terms of information content. For the purposes of indexing and summarizing the news, national and world leaders, terrorists, war criminals, and corporate leaders are key entities who tend to occur frequently in nearly all news sources over a certain period of time. New names in this category constantly are becoming newsworthy, and these names tend to remain in the news. However, it is usually impossible to anticipate new names of global importance that will enter the news. We can use a technique similar to the [38] method to provide candidates for the initial name list, assuming that names of global importance would appear in both broadcast and print news sources during the same period.

News reporters. The names of news anchors and reporters are very common in broadcast news, since most story segments begin and end with a "hand-off" to or from a reporter (such as "CNN's John Zarrella has the story..."). While these names may not be of global importance, they cause a large number of OOV errors that can also cause additional errors in an utterance. Fortunately, the names of anchors and reporters for a specific broadcast news source are usually readily available, either from the news agency itself or from manual transcripts of other broadcasts.

Spelling and Morphological Variants. Many instances of out-of-vocabulary words are actually very similar to words in the vocabulary, but differ slightly due to spelling or morphological variations. As mentioned earlier, the words "Coburns" and "Lewinsky" are OOV, although "Coburn" and "Lewinski" are in the lexicon. The names "Dench" and "Jennings" are OOV because of spelling errors in the transcription. Finally, there are some commonly used orthographic variants for foreign names (e.g. Gadhafi, Qadafi, and Khadafi).

Sports Figures. Many news broadcasts provide a variety of different segments, including not only world, national, and local news, but also business, sports and weather reports. Consequently, one of the primary sources of OOV words is names of sports figures mentioned in sports reports. While the number of athletes that might potentially be mentioned on the news is as unlimited as the number of new names of global

[2] Some examples of these rare surnames are Barrymore, Namath, Travolta, Garciaparra, Stephanopoulos, Nimoy, and Chomsky.

OOV Name Category	List Source	List Size	Token Coverage	Type Coverage
Global names	NYT/APW + Census	199,825	71%	63%
News reporters	Broadcast agency	140	95%	98%
Spelling/morphology	Census Homophones	23,615	92%	94%
Sports figures	Rosters	1,300	94%	96%
Villagers	NYT/APW + Census	199,825	21%	14%
All	All	201,130	61.5	52.0

Table 2: *Major categories of OOV names and lexical resources that provide good coverage for the categories.*

importance, we can provide excellent vocabulary coverage of this category by focusing on the major sports that might be discussed on the broadcasts in the corpus. In our case, we can achieve nearly perfect coverage of sports names by including the active rosters from the baseball (MLB), football (NFL), basketball (NBA), and hockey (NHL) leagues, as well as the top 100 rated tennis and golf professionals on the major tours.

Villagers and Human Interest Personalities. Perhaps the most difficult category of OOV names is “incidental” names that are mentioned in the course of a news story or human interest story. These OOV names usually correspond to residents of a town hit by a newsworthy natural disaster (e.g., “Cyrofin Crevin has been living in Dzershinsk since 1939”) or to participants in events of very limited interest, like the 1831 steamboat race on the Hudson River. In many cases, the names are very specific to the geographic area in which the story takes place, and good vocabulary coverage would require lexical resources particular to the language common in the region of interest. Coverage of these names is very poor using any of the resources applicable to the other categories. However, in practice, this category of names is the least important in terms of information content; for indexing and summarization purposes the incidental names are much less important than the major figures, like heads of state and government officials.

Table 2 provides a summary of the major OOV name categories, with the word token and word type coverage for the lexical resource best matched to that category. In all, 61.5% of the OOV tokens and 52.0% of the OOV types in the training set were contained in at least one of the lexical resources, which is significantly better than the Census name lists alone (31.4% token/29.1% type). This suggests that few of the OOV names in the news are common American surnames and that the task-dependent name vocabularies are a better fit to the news broadcast than the extensive, generic list of names.

5.2 Filtering Vocabulary Lists Using Text-Based IE

The combined lexical resources described in the previous section provide excellent coverage for most categories of OOV names. However, the NYT/APW list consists of over 150,000 different words alone and 200,000 when combined with the Census list. Instead of using all words from the NYT/APW data, we would prefer to consider only the words that are likely to be contained in name phrases. One option for doing this would be to manually filter the NYT/APW word lists to remove all words that are not likely to be names, which of course would be a very time-consuming process. An alternative to manual filtering is to use a text-based IE system to automatically label person phrases.

In this work, we use a statistical IE system trained to identify names of persons, places and organizations. It represents names using hidden states, which are associated with state-dependent word bigrams and word-dependent state transitions [28]. Name extraction systems are typically evaluated in terms of the F-measure,

Lexical	List	Overall	Overall	CNN	CNN	ABC	ABC
Resource	Size	Token	Type	Token	Type	Token	Type
Census	93,963	31.4	29.1	22.8	26.2	13.1	27.3
APW (all words)	94,289	29.4	21.7	35.8	29.2	77.6	41.8
NYT (all words)	112,345	41.7	31.0	52.5	46.2	85.6	60.0
NYT/APW (all)	162,811	48.6	38.8	53.1	47.7	89.0	70.9
Census+NYT/APW (all)	201,130	61.5	52.0	66.0	60.0	94.2	78.2
APW (IE names)	15,475	25.7	17.0	34.0	26.2	76.1	36.4
NYT (IE names)	22,537	37.5	25.1	48.1	40.0	83.7	52.7
NYT/APW (IE names)	33,244	43.5	32.1	48.8	41.5	87.1	63.6
Census+NYT/APW (IE)	102,559	57.1	46.4	62.3	55.4	92.3	70.9

Table 3: *Coverage of OOV name words in training data for various lexical resources, including name lists automatically generated by our text-based IE system.*

which is the harmonic mean of recall and precision rates. Our IE system has a performance of about F=90 on text data, and it can process the entire NYT/APW corpus in minutes.

Table 3 shows the effect on vocabulary coverage of limiting the candidate list to the person words automatically labeled in the NYT/APW texts by our text-based IE system. Using the IE system provides an 80% reduction in the candidate list size, from 162,811 to 33,244 words, while degrading the coverage by only about 5%.

Table 3 also shows coverage results for the two broadcast sets, CNN Headline News and ABC World News from January 1998, that directly overlap with the NYT/APW texts. It is interesting to note that the CNN results are almost identical to the overall result, while the ABC results are significantly better. In fact, the NYT/APW word lists provide nearly 90% coverage of the ABC broadcasts, and there is almost no degradation in using the automatically-generated person list. This highlights the differences between broadcast types in the corpus and illustrates the fact that vocabulary coverage is very much source dependent.

It is also interesting to note that the word token coverage is often significantly higher than the word type coverage in Table 3. This can be explained in part by the "bursty" nature of news, in which certain names tend to be mentioned very frequently over a period of time, once they become newsworthy. For example, of 326 OOV person tokens in the ABC News data, 200 (61%) were "Lewinsky," a name present in the news very frequently in January 1998. This one word therefore represents only 1.8% of the OOV types but 61% of the OOV tokens; in contrast, the OOV word "Shiong" occurs only once in the ABC data and also represents 1.8% of the OOV types but only 0.3% of the OOV tokens. This can account for some of the large difference between the ABC results and the overall (and CNN) results. The word type coverage (which only counts "Lewinsky" one time) is still much higher for the ABC data than the other sources, however, indicating that there is a significant source difference. It may simply be that ABC ran mainly high interest stories during this period that overlapped with other news sources, like the New York Times and AP Newswire, while CNN broadcast many unique stories that did not overlap with other sources.

The inherent tradeoff in list size vs. OOV coverage is shown in Figure 3, which provides a graphical representation of the percent of OOV words contained in name lists of various sizes. The graph shows the coverage as the list size increases for all OOV words in the training set and for OOV words in the ABC News broadcasts. The two lower curves show the coverage of all OOVs and ABC OOVs using subsets of various sizes from the lists of all words from the NYT/APW texts. The coverage of the ABC OOVs is clearly significantly higher than for the full training set. The second set of curves shows the coverage of

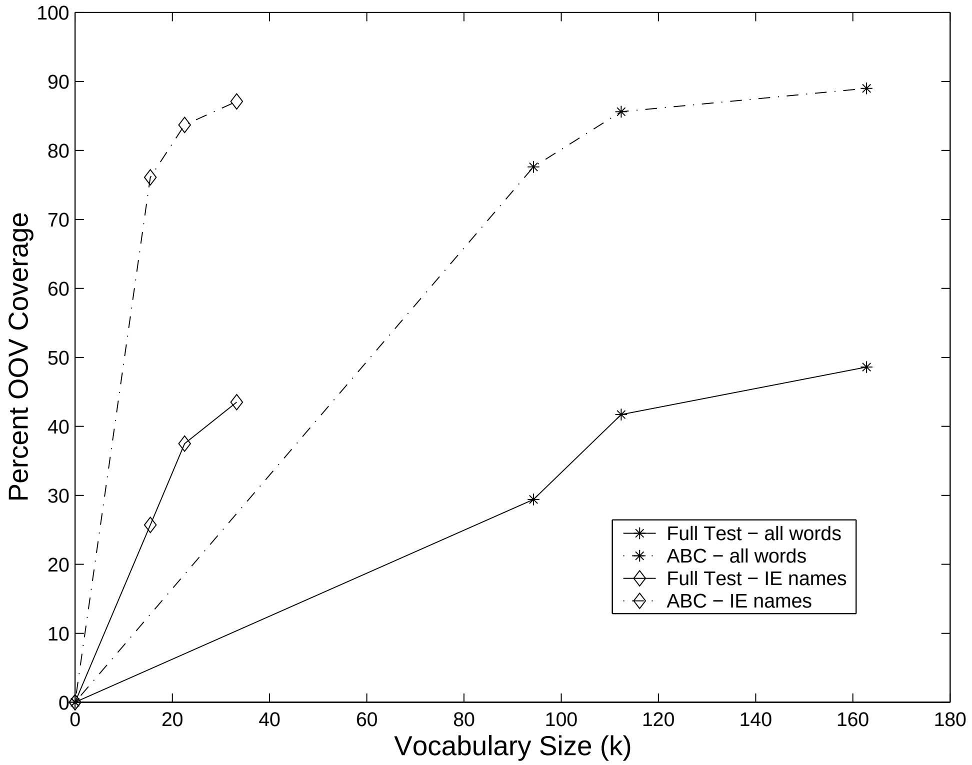

Figure 3: Tradeoff in list size vs. OOV coverage for entire training set and for ABC News subset.

all OOVs and ABC OOVs using lists of person words automatically identified in the NYT/APW texts by our text-based IE system. There is a striking difference in the list sizes, yet the coverage provided is nearly identical. Alternatively, we can see that for a fixed list size, the coverage can be doubled or tripled for the full test set, and much more for the ABC subset.

This method for name list generation and filtering provides an excellent source of candidates for correcting OOV ASR errors. However, the list is quite long, and the presence of an OOV word in a candidate list does not guarantee that we can identify it as the most likely replacement. In the next section, we discuss how we can use a phonetic distance to prune the list, and optionally to correct the OOV errors.

6 Phonetic Distance and List Ranking

In this section we describe a novel technique that treats ASR error correction as a form of spelling correction similar to the noisy channel approach (described in Section 3.2). We measure the edit distance between two sequences of characters (phonemes) according to a trainable weighting system, and rank candidate corrections according to their phonetic distance.

6.1 Phonetic Distance

Differences between phonemes in two sequences are defined according to a weight matrix that determines the cost of substituting, inserting, or deleting a phoneme. One option would be to use a phonetic feature-based weighting function, as in [3, 30]. However, in experiments with different alternatives, we found that the best results were obtained with automatically derived weights. Hence, the calculation of the phonetic distance between two words (or sequences of words) consists of two components: offline learning of the weights, and online string alignment.

We use a set of weights derived empirically from training data using the EM technique developed in [32], as described in Section 3.2. In our weight estimation, we use a set of ASR output from a portion of the TDT data separate from that used in our other experiments. The ASR words are automatically aligned with the reference transcriptions from the TDT data, and then both the reference words and the ASR output words are automatically converted to sequences of phonemes using a text-to-phoneme program.[3] In the experiments reported here, we use the t2p program [22]. In essence, we treat the errorful ASR output as a set of phonemic misspellings that need to be transduced into the human-generated reference transcripts. We then run the weight estimation algorithm on the phonemic word pairs to compile stochastic weights for each possible phonemic insertion, deletion, and substitution.

The learned weights are applied according to the best alignment between the phonemes in the two strings, determined by an efficient dynamic programming search of possible phoneme alignments.

6.2 Distance-Based Name List Pruning

The phonetic distance framework can be used to further prune the vocabulary lists generated by the text-based IE system in Section 5.2, eliminating all candidates that are not close enough phonetically to the hypothesized word or words. To accomplish this distance-based pruning, we simply calculate the phonetic distance between the ASR hypothesis and each of the words in the candidate list. We then sort the candidate list from shortest distance to longest and determine where the OOV reference word is ranked. ¿From these rankings for all OOV tokens in the corpus, we can determine the relationship between coverage and pruned list size.

Figure 4 demonstrates the significant benefit provided by using the phonetic distance calculation to rank the lists of candidate words. The pairs of curves represent the results for OOV coverage from Figure 3 for the full set of OOV persons in the training set (the lower pair of curves) and for the ABC portion of the training set (the upper pair). The first curve of each pair is the coverage for the list of person names identified by the text-based IE system in the NYT/APW data; the second curve is the coverage for the same list after ranking according to the phonetic distance. In both cases, the full training set and the ABC subset, pruning the list according to phonetic distance ranking allows for nearly identical coverage with a list size of just 5,000 words, an 85% reduction over the 33,000 word list.

6.3 Phonetic Distance for ASR Error Correction

The phonetic distance technique described in this paper is a very effective tool for ranking and pruning the candidate name list. In fact, the ranked list also can be used directly to correct ASR errors within name phrases by considering the top-ranked candidate to be the replacement for the hypothesized word if the distance is sufficiently small. We evaluated this error correction method in two ways, both in terms of the number of errors corrected and in terms of the impact on information extraction. Since the error token detection algorithm does not have high accuracy, we used the error correction in conjunction with both automatic and oracle (perfect) name and OOV token detection, to factor out the effect of OOV detection errors.

In all experiments, we used the pruned 33,000 word person list produced by the text-based IE system from the NYT/APW data. Since the vocabulary coverage of this resource was 43%, that was the maximum number of the OOV tokens that we could possibly correct. Table 4 shows that the phonetic distance calculation corrected 16.4% of the OOV person names, which represents nearly 40% of those names covered by

[3] While it would be more efficient to use the original ASR pronunciation dictionary, it was not available for this work and the use of grapheme-to-phoneme prediction for both IV and OOV words has the possible advantage of consistency.

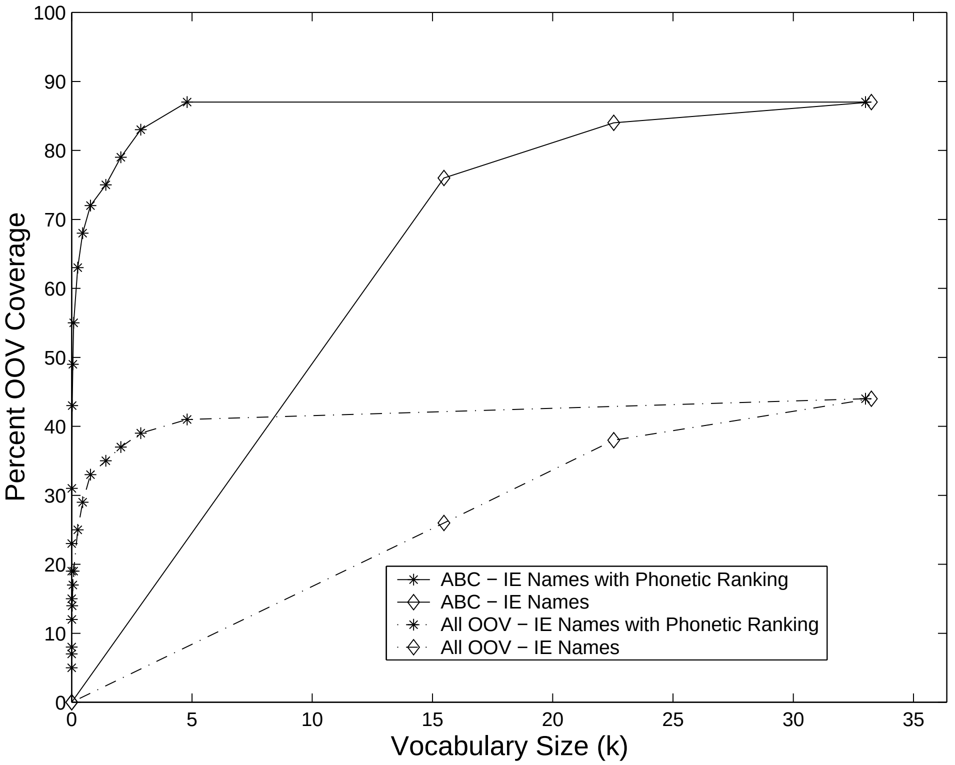

Figure 4: Effect on OOV coverage of pruning phonetically-ranked lists, for entire training set and for ABC News subset.

IE and Correction	% Corrected	
Method	OOV errors	IV errors
Oracle IE, Known OOV errors	16.4	0
Auto IE, Detected OOV errors	1.1	3.6

Table 4: *Error correction results from using top candidate in phonetically-ranked list.*

the name list. For the fully automatic case, where there may be some name detection and OOV detection errors, there are also some in-vocabulary words corrected.

While correcting ASR errors is often a goal in itself, since it results in improved overall word error rates for the speech recognizer, our work in error correction has a slightly different objective. We focus on error correction in regions of speech that have high information content, specifically names, which impacts the quality of the information that is extracted from the ASR output. For IE of name phrases in speech data, the standard evaluation measure accounts for type (person, place, etc.), extent and content errors by averaging the F-measure scores associated with each error type [8]. Since content errors are due entirely to the ASR system, the content F-measure can actually decrease when the extent detection is improved because of OOV names. Therefore, OOV name correction should translate primarily to an improvement in the content score, with a smaller gain in the overall score because of score averaging. Table 5 shows that correcting 16% of the OOV names resulted in an improvement of the content score from F=64.8 to F=66.9 in the oracle system, and an improvement in the overall IE score from F=86.4 to F=87.2. For the fully automatic system, even a small number of corrections impact the content score, which improves from F=66.9 to F=68.4. Note that

IE and Correction Method	Name Content	Name Overall
Auto IE, No Correction	66.9	73.4
Auto IE, Detected OOV errors	68.4	74.0
Oracle IE, No Correction	64.8	86.4
Oracle IE, Known OOV errors	66.9	87.2

Table 5: *Improvements in name phrase extraction performance (F-measure) associated with error correction, based on using top candidate in phonetically-ranked list.*

Britany, Britney, Britni, Brittaney, Brittani, Brittanie, Brittany, Britteny, Brittney, Brittni, Brittny
Caal, Coll, Colle, Cull, Kahle, Kall, Kol, Kole, Koll, Kolle, Kull
Eurich, Urich, Urick, Urik, Yurich, Yurick
Hsu, Hsueh, Seu, Soo, Sou, Su, Sue, Tsou, Tsu
Laurey, Laurie, Laury, Lawrey, Lawrie, Lawry, Loree, Lorey, Lori, Lory
Wolf, Wolfe, Wolff, Wolffe, Woolf, Wulf, Wulff

Table 6: *Examples of normalized names identified by phonetic distance calculation in the census list.*

while this error correction has a direct impact on information extraction, it has an insignificant effect on the overall word error rate of the data, since the number of words corrected are a small percentage of the total.

6.4 Name Normalization

As we described in Section 5.1, one of the main sources of OOV names in text is alternate spellings and morphological variants of in-vocabulary words. The phonetic distance framework can help us identify sets of homophones, words that are spelled differently but pronounced the same. Using these homophone sets constructed for target names, we can determine whether a word that is technically out-of-vocabulary can actually be considered equivalent to an in-vocabulary word, thus providing a straightforward method for name normalization.

We used the phonetic distance framework to identify all sets of words in the Census name list with identical pronunciations (a phonetic distance of zero), again using the pronunciations generated by the t2p program [22]. From the list of 93,963 orthographically distinct names, we identified 9,638 sets of homophones containing 23,615 names, thus reducing the list to just 79,986 phonetically distinct names. Table 6 shows some examples of the normalized name sets identified. One particularly interesting aspect of the automatically-generated homophone sets is the broad range of orthographic differences that can be normalized.

7 Conclusions and Future Work

Since new words are constantly being introduced into common usage, it is impossible to ever have a complete vocabulary of all spoken words. The treatment of new lexical items is thus an essential element of any system aiming to process natural language. In this paper, we focused particularly on the problem of recognizing new names in speech, showing with an analysis of the TDT data that names are the dominant source of out-of-vocabulary words for an ASR system designed to recognize spoken news broadcasts. We describe a framework for correction of OOVs, involving four key components: offline name list generation, detection of candidate OOV errors, online list pruning based on phonetic distance, and final candidate selection. For offline name list generation, our analysis identified the most important name types and lexical resources that give high coverage of some categories, as well as presented a text filtering technique using automatic name phrase extraction on temporally relevant texts to identify new names of global importance or associated with major news events. We also showed that the name list could be pruned by nearly an order of magnitude in an online stage of computing the phonetic distance of candidate name pronunciations to the phone sequence associated with the target OOV region in the ASR output. In experiments on OOV correction, we showed that 40% of the covered OOV words could be corrected by a phonetic distance alone with perfect OOV detection, and that accurate OOV detection is the major limiting factor at this point. Even with poor OOV detection performance, however, the correction algorithm still results in improvement to automatic name phrase extraction performance on speech data. Finally, in a pilot study, we showed that the same name list generation and phonetic distance ideas are useful for generating homophone lists, which has implications for name normalization in language processing.

There were several simplifying system design choices made in the OOV correction system that could be changed to obtain improved performance. Since the time of this work, significant advances have been made using confusion networks in word confidence estimation [12], which would likely lead to higher accuracy OOV detection. In name list generation, we did not prune the census or other list resources by frequency, and it may be that only the most frequent names are important given the additional use of temporally relevant text. The phonetic distance ranking step could be improved by having richer output from the ASR system, e.g. a phoneme lattice that could be obtained from a word lattice or confusion network. In addition, the work here was limited to correction of single words, but performance would be much improved if multi-word sequences were modeled because the ASR output words are frequently not one-to-one with the target words. For example, the OOV inputs “Orla Guerin” resulted in the ASR output “*or legere and*.” According to the automated alignment, which allowed only one-to-one word alignments, “Orla” aligned with “*or legere*,” and “Guerin” aligned with “*and*.” Needless to say, our phonetic distance ranking produced poor results for these examples using a single word distance, d(“*or legere*”, “Orla”) and d(“*and*”, “Guerin”). Our method would have ranked the correct answer quite high if the two sequences were used e.g. d(“*or legere and*”, “Orla Guerin”). Lastly, the final ranking stage used no language model scores, which are clearly very important, particularly for homophones or spelling variants that one would want to distinguish between. Since the words were not in the original vocabulary, the original language model provides no information for this task. However, one can use the candidate name list to collect additional text data for adapting the language model, building on recent work such as [35, 7]. An additional improvement in our phonetic distance calculation might be achieved by using the richer string edit syntax developed in [6] rather than the restricted one-to-one character edits we use.

While our focus was on English broadcast news, where names are an important problem, for other languages and/or other tasks, the biggest problem may be dealing with morphological variants. Typically, an ASR vocabulary must explicitly contain the exact form of a word for that variant to be output, e.g. plural and possessive forms in addition to the singular form. This results in many errors in which the OOV variant is spoken (“Kelly’s” or “Kellys”), but the in-vocabulary base form is output (“Kelly” or “Kelly is”). This problem can also be addressed in our framework because of the flexibility in the weight matrix, by assigning a weight of zero to word-final insertion and deletion of the phonemes S or Z (and possibly also a preceding vowel like AX). Other morphological variants, such as inflectional verb endings, can also be addressed using this technique. We can thus create morphological equivalence classes similar to the homophone sets, and use class language models to select among the alternatives.

The homophone lists we generated from the extensive Census name lists can be used in many natural language processing applications. For example, a major problem in information retrieval (IR) is accurately and efficiently indexing spelling and morphological variants. An IR system essentially takes a set of training documents and creates an index of which words occur in which documents, how frequently they occur in those documents, and which other words they occur with. If indexed words have multiple spelling variants (e.g., Qadafi and Khadafi), each variant will be indexed separately and will thus be difficult to later retrieve together, even though they represent the same entity. Our sets of equivalent names can help disambiguate misspellings and alternate spellings of indexed terms. In this way, if equivalence classes are indexed by the IR engine, documents containing "Qadafi" would be returned when a variant such as "Khadafi" is contained in the query. Note that contextual language cues will typically be needed to determine when different variants actually refer to the same entity. In addition, when a word has multiple pronunciations, then it will be assigned to more than one equivalence class, and again other language cues can help refine selection to the proper documents.

The problem of name resolution is of growing importance, as improved performance in speech and language technology has made large-scale automatic processing of spoken documents feasible. Out-of-vocabulary word handling is a key limitation of current technology. The work presented here provides a general framework for new word (and particularly new name) handling in language processing and a specific solution for word list generation components, but research leveraging these advances in other components would lead to further performance gains. In particular, exploring connections between OOV detection and adaptive language modeling is one important next research direction.

Acknowledgments. A portion of this work was completed while the first author was affiliated with the University of Washington and the MITRE Corporation. The weights we use in the phonetic distance experiments were derived by John Henderson, and the alignment tool was developed by Chad McHenry. We thank them for making these resources available for our work.

References

[1] D. Appelt and D. Martin, "Named Entity Extraction from Speech: Approach and Results Using the TextPro System," *Proc. DARPA Broadcast News Workshop*, pp. 51-54, 1999.

[2] A. Asadi, R. Schwartz, and J. Makhoul, "Automatic Detection of New Words in a Large Vocabulary Continuous Speech Recognition System," *Proc. International Conference on Acoustic, Speech and Signal Processing,* pp. 125-128, 1990.

[3] R. Bates and M. Ostendorf, "Reducing the Effects of Pronunciation Variability on Spontaneous Speech Recognition using Prosody and Discourse," *Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding,* pp. 17-22, October 2001.

[4] I. Bazzi and J. Glass, "Modeling Out-of-Vocabulary Words for Robust Speech Recognition," *Proc. International Conference on Spoken Language Processing,* vol. I, pp. 401-404, 2000.

[5] M. Boros, M. Aretoulaki, F. Gallitz, E. Nöth, and H. Niemann, "Semantic processing of out-of-vocabulary words in a spoken dialogue system," *Proc. European Conference on Speech Comm. and Tech.,* vol. 4, pp. 1887-1890, 1997.

[6] E. Brill and R. Moore, "An Improved Error Model for Noisy Channel Spelling Correction," In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics* (ACL00), pp. 286-293, Hong Kong, 2000.

[7] I. Bulyko, M. Ostendorf, and A. Stolcke, “Class-dependent interpolation for estimating language models from multiple text sources,” UWEE Technical Report, no. UWEETR-2003-0003, 2003.

[8] J. Burger, D. Palmer, and L. Hirschman, “Named Entity Scoring for Speech Input,” *Proc. Annual Meeting of the Association for Computational Linguistics (COLING)*, pp. 201-205, August 1998.

[9] G. Chung, “Automatically Incorporating Unknown Words in Jupiter,” *Proc. International Conference on Spoken Language Processing*, vol. IV, pp. 520-523, 2000.

[10] K. Church and W. Gale, “Probability Scoring for Spelling Correction,” *Statistics and Computing*, 1:93–103, 1991.

[11] A. Dempster, N. Laird, and D. Rubin, ”Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society (B)*, 39(1):1-38, 1977.

[12] G. Evermann and P.C. Woodland, “Posterior Probability Decoding, Confidence Estimation and System Combination,” *Proc. of the NIST Speech Transcription Workshop*, University of Maryland, 2000. `http://www.nist.gov/speech/publications/tw00/`

[13] P. Fetter, “Detection and Transcription of OOV Words,” Verbmobil Technical Report 231, August 1998.

[14] P. Geutner, M. Finke, and P. Scheytt, “Adaptive Vocabularies for Transcribing Multilingual Broadcast News,” *Proc. International Conference on Acoustic, Speech and Signal Processing*, vol. II, pp. 925-928, 1998.

[15] P. Geutner, M. Finke, and A. Waibel, “Phonetic-Distance-Based Hypothesis Driven Lexical Adaptation For Transcribing Multlingual Broadcast News,” *Proc. International Conference on Spoken Language Processing*, vol. 6, pp. 2645-2638, 1998.

[16] L. Gillick, Y. Ito, and J. Young, “A Probabilistic Approach to Confidence Estimation and Evaluation,” *Proc. International Conference on Acoustic, Speech and Signal Processing*, vol. 2, pp. 879-882, 1997.

[17] L. Gillick, Y. Ito, L. Manganaro, M. Newman, F. Scattone, S. Wegmann, J. Yamron, and P. Zhan, “Dragon Systems’ Automatic Transcription of New TDT Corpus,” *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pp. 219-222, 1998.

[18] I. L. Hetherington, *A Characterization of the Problem of New, Out-of-Vocabulary Words in Continuous-Speech Recognition and Understanding*, PhD Thesis, Massachusetts Institute of Technology, 1995.

[19] R. Iyer and M. Ostendorf, “Transforming Out-of-Domain Estimates to Improve In-Domain Language Models,” *Proc. European Conference on Speech Comm. and Tech.*, vol. 4, pp. 1975-1978, 1997.

[20] D. Jurafsky and J. Martin, *Speech and Language Processing*, Prentice Hall, 2000.

[21] M. Kernighan, K. Church, and W. Gale, “A Spelling Correction Program Based on a Noisy Channel Model,” *Proc. Annual Meeting of the Association for Computational Linguistics (COLING)*, Vol. II, pp. 205-211, Helsinki, Finland, 1990.

[22] K. Lenzo, “s/($text)/speech $1/eg;” *The Perl Journal*, vol. 3, No. 4 (#12), Winter 1998.

[23] V. I. Levenshtein, “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals,” *Soviet Physics Doklady*, 10:707-710, 1966.

[24] B. Logan and J. M. Van Thong, "Confusion-based query expansion for OOV words in spoken document retrieval," *Proc. International Conference on Spoken Language Processing,* vol. 3, pp. 1997-2000, 2002.

[25] E. Mays, F. Damerau, and R. Mercer, "Context Based Spelling Correction," *Information Processing and Management*, 27(5):517-522, 1991.

[26] D. Palmer and M. Ostendorf, "Improving Information Extraction by Modeling Errors in ASR Output" *Proc. Human Language Technology Workshop,* pp. 156-160, 2001.

[27] D. Palmer and M. Ostendorf, "Improved Word Confidence Estimation using Long Range Features," *Proc. European Conference on Speech Comm. and Tech.,* vol. 3, pp. 2117-2120, 2001.

[28] D. Palmer, *Modeling Uncertainty for Information Extraction From Speech Data,* PhD Thesis, University of Washington, 2001.

[29] D. Palmer, M. Ostendorf, and J. Burger, "Robust Information Extraction from Automatically Generated Speech Transcriptions," *Speech Communication*, vol. 32, pp. 95-109, 2000.

[30] M. Riley and A. Ljolje, "Automatic generation of detailed pronunciation lexicons," in *Automatic Speech and Speaker Recognition: Advanced Topics,* ed. C.-H. Lee, F. K. Soong and K. K. Paliwal, Kluwer, Boston, 1995.

[31] E. Ringger and J. Allen, " Error Correction via a Post-Processor for Continuous Speech Recognition," *Proc. International Conference on Acoustic, Speech and Signal Processing,* vol. I, pp. 427-430, 1996.

[32] E. Ristad and P. Yianilos, "Learning String Edit Distance," *IEEE Trans. PAMI,* 20(5):522-532, 1998.

[33] R. Rosenfeld, "Optimizing Lexical and N-gram Coverage via Judicious Use of Linguistic Data," *Proc. European Conference on Speech Comm. and Tech.,* vol. 2, pp. 1763-1766, 1995.

[34] T. Schaaf, "Detection of OOV words using generalized word models and a semantic class language model," *Proc. European Conference on Speech Comm. and Tech.,* vol. 4, pp. 2581-2584, 2001.

[35] S. Schwarm and M. Ostendorf. "Text normalization with varied data sources for conversational speech language modeling," *Proc. International Conference on Acoustic, Speech and Signal Processing,* vol. I, pp. 789-792, 2002.

[36] B. Suhm, M. Woszczyna, and A. Waibel, "Detection and Transcription of New Words," *Proc. European Conference on Speech Comm. and Tech.,* vol. 3, pp. 2179-2182, 1993.

[37] United States Census Bureau Name Files, (www.census.gov/genealogy/names), 1995.

[38] H.D. Wactlar, T. Kanade, M.A, Smith, and S.M. Stevens, "Intelligent Access to Digital Video: Informedia Project," *IEEE Computer,* 29(5):46-52, 1996.