INTERSPEECH 2009 BRIGHTON

Improving the Recognition of Names by Document-Level Clustering

Bin Zhang[1], *Wei Wu*[1], *Jeremy G. Kahn*[2], *Mari Ostendorf*[1,2]

Signal, Speech and Language Interpretation Lab
[1]Department of Electrical Engineering, [2]Department of Linguistics
University of Washington, Seattle, WA 98195, USA
{binz,weiwu,jgk,ostendor}@u.washington.edu

Abstract

Named entities are of great importance in spoken document processing, but speech recognizers often get them wrong because they are infrequent. A name correction method based on document-level name clustering is proposed in this paper, consisting of three components: named entity detection, name clustering, and name hypothesis selection. We compare the performance of this method to oracle conditions and show that the oracle gain is a 23% reduction in name character error for Mandarin and the automatic approach achieves about 20% of that.
Index Terms: ASR, name, clustering, confidence

10.21437/Interspeech.2009-319

1. Introduction

Named entities (NEs) – names of people, places, and organizations – are important to applications including machine translation and information extraction. When these applications take input data from the automatic speech recognition (ASR) output, the name errors in the ASR output can degrade their performance. Therefore, improving the ASR of names, has potential benefit to downstream applications. According to our statistics, using a state-of-the-art Mandarin speech recognizer [1, 2], most of the NE recognition errors are in person names, and the character error rate (CER) in the person name regions is two to three times higher than the overall CER. This paper is thereby focused on reducing the recognition error of person names in the context of a Mandarin speech recognizer. For succinctness, we will use "name" to refer to "person name" hereafter.

Conventional speech recognizers are trained to optimize the overall recognition accuracy, and no special treatment for NEs is employed. All words, including names and non-names, are dealt with equally. However, names are different from non-names in certain respects. For example, unlike common words, some names in the test documents may not be seen in the training documents, thus they are likely to be out-of-vocabulary (OOV) words for the speech recognizer. In the situation of Mandarin ASR, these names can be broken down into individual characters. In many cases, they are recognized as strings of characters that are different from the actual ones in the names, some of which do not look like names at all. Consequently, the NE detector may miss them, and the downstream applications will be affected. Motivated by the importance of names, algorithms to deal with OOV names in English ASR were proposed in [3], where OOV names were detected, resolved and corrected according to name lists generated off-line. A limitation of this work was in the low recall of OOV detection. Liu *et al.* [4] proposed using NE information for topic analysis and language model adaptation. The target, however, was still to reduce the overall recognition error rate. Recently in [5], a name-aware speech recognizer for interactive question answering system was proposed. This system built an NE-specific language model from the document containing the user-provided NE. Substantial improvements in word error rate were achieved. However, the user needs to provide a target NE before asking the question, which is not possible in many applications. In this paper, we present a method that corrects the name recognition error in terms of a post-processing to the ASR output. This method is composed of three components: automatic NE detection in ASR output, document-level name clustering, and cluster-wise name correction by selecting the name hypothesis with the highest total confidence over multiple mentions.

Although the characters from the ASR output in the name region may not be trustworthy, the name can still be characterized by acoustic cues, such as acoustic features and phones. Ji *et al.* [6] proposed using phones, rather than characters, from the ASR output to perform cross-lingual spoken sentence retrieval. The query names and document were converted to a phonetic representation, and fuzzy name matching confidence estimation was calculated based on a phonetic string distance. Hence, some of the names that are incorrectly recognized by the speech recognizer can be retrieved. In our work, a phonetic distance of names is also used. A spectral clustering [7] algorithm is built on top of the phonetic distance. Several experiments with different oracle settings on each component were conducted to assess the limitation of each processing step. Improvements on ASR name errors were obtained, which approaches the limit established by oracle name clustering and oracle name hypothesis selection in the context of automatic NE detection.

This paper is organized as follows. The details of each step of our name correction algorithm are introduced in section 2. In section 3, we present our experimental data, oracle experiment set-up, experimental results, and some extensive comparison and discussion of our results. Section 4 concludes the paper and suggests some future directions.

2. Method

When humans hear a name that they have never heard (e.g. in a news story), they might not get the name correct at the first time. However, as the same name is heard again and again in other phonetic contexts, the listener will have the chance to recognize it correctly. Afterwards, the correct name is memorized and all the previous recognitions, even if previously understood differently, are understood to refer to the same (now correct) name. This intuition about human behavior motivated our method to improve the ASR of names.

In this work, we implemented an algorithm to correct the name errors in the ASR output. The method we used includes the following steps:

1. Automatically detect the names in the ASR output of a document.
2. Cluster these names based on their acoustic similarity.
3. Force the names in a cluster to be consistent using confidence-weighted voting.

The three major components that correspond to these steps will be discussed in detail in the following sections.

2.1. Named entity detection

The NE tags are from the output of the Mandarin NE detection system described in [8]. It is built on a hidden Markov model NE tagger, and it also uses other information in rescoring, including co-reference, relation, mention and semantic role. It also provides English translations of the names. This NE detection system has good performance on text data. On ASR output, however, the performance drops due to the ASR errors. This system is capable of detecting multiple types of NEs including person, geo-political entity, organization and location. Only the tags of the person type are used in this paper.

2.2. Document-level name clustering

Our hypothesis is that names with sufficient acoustic similarity in the same document should all refer to the same person. Based on this assumption, all the names detected in a document are clustered according to acoustic similarities. Since name hypotheses have different lengths, we represent the names in a document in terms of their pairwise distances rather than as vector elements. Because of the representation, we use graph-based rather than centroid-based clustering.

The distance is based on the phone sequence, instead of directly on acoustic features,[1] to normalize out variation across speakers and accents. A phonetic distance matrix is created based on the phonetic distances of all the pairs of names in a document. Given two phone strings p_i and p_j of two names n_i and n_j, respectively, to compute the phonetic distance d_{ij}, we use the Levenshtein edit distance between the two phone strings via dynamic programming. A weighted edit distance based on a phonetic confusion matrix [9] did not seem to improve the performance.

Given a distance matrix, a fully connected similarity graph with the nodes being the name phone strings can be constructed. The similarity between nodes p_i, p_j is computed by mapping the distances to real values between 0 and 1 using an exponential function

$$s(p_i, p_j) = \exp(-d_{ij}). \tag{1}$$

A standard approach to graph-based clustering is spectral clustering [7]. The spectrum of the graph reveals the underlying clusters. The procedure of this clustering algorithm is as follows. First, the eigenvalues of the graph Laplacian are extracted, and an $n \times k$ matrix V containing the first k eigenvectors is obtained, where n is the number of nodes, and k is the predicted number of clusters. Next, the rows of V are normalized to form a new matrix U with the same dimension [10]:

$$u_{ij} = \frac{v_{ij}}{\sqrt{\sum_k v_{ik}^2}}. \tag{2}$$

Finally, a bottom-up agglomerative clustering is carried out on the rows of U.

[1] We tried computing the similarity of cepstral feature sequences of words using dynamic time warping, but performance was worse than for phones, particularly in determining the correct number of clusters.

In general, the determination of the number of clusters k is not easy. Researchers have used the "eigen gap" concept [7] to locate the first large gap in eigenvalues which indicates the number of clusters. However, it did not work well in our experiments. We count the eigenvalues with magnitude below a threshold θ to set k in this paper. θ is adjusted to yield the smallest global clustering error on a development set. This simple determination of the number of clusters also makes spectral clustering attractive in our application.

2.3. Name hypothesis selection

Once the membership of each name to each cluster is determined, we enforce consistency of names in clusters with more than one mention. Since in real applications, the correctness of the name hypotheses is unknown, we choose the best name hypothesis using a confidence measure-based majority voting.

In the ASR output, each word is assigned a confidence measure. It is computed based on the geometric mean of character-level confusion network posteriors. The confidence measure of word w_i can be approximately considered as the probability of w_i being correctly recognized. Assuming there are m name hypotheses $N = \{n_1, \ldots, n_m\}$ in a name cluster, and the unique elements in N form the set $\hat{N}$, the best name hypothesis is picked using

$$n^* = \underset{n \in \hat{N}}{\operatorname{argmax}} \sum_{i:\ n_i = n} c_i, \tag{3}$$

where $c_i, i = 1, \ldots, m$, are the confidence measures of the names.

3. Experiments

In order to get good performance in name correction, all the three components described above should ideally achieve satisfactory performance. However, this is not exactly the case. For example, the NE detection system itself is very complicated, and NE detection on errorful ASR output, which does not have very good accuracy, is still an open research topic. The performance concern in clustering and confidence measure estimation is not as big as in NE detection, but any inaccuracy will inevitably degrade the overall performance of name correction. According to our statistics, the equal error rate of the ASR confidence measure on the data used in this experiment is about 20%. To find out more about where the performance limit of this method is, we conducted experiments with different combinations of "automatic" vs. "oracle" conditions for each component. The oracle components are defined as follows.

Oracle NE detection This NE detection is based on human labels. The ASR reference is labeled by human in advance, and is aligned to the ASR hypothesis. The oracle NE detector simply looks up the ASR hypothesis and the alignments, and returns the name hypotheses which are aligned to the labeled names in the ASR reference.

Oracle name clustering This name clustering in the ASR hypothesis is simply the clustering of the names based on the reference strings which they are aligned to. That is to say, in a document, the names in the ASR hypothesis with the same corresponding reference strings are put in the same cluster.

Oracle name hypothesis selection As we have aligned the ASR hypothesis with the ASR reference, the number of

character errors of each name hypothesis can be computed. The best name hypothesis in a cluster is then chosen to be the one with the fewest errors.

3.1. Data

We used the Mandarin ASR output from the dev, test, and eval sets of the Nightingale team in the Global Autonomous Language Exploitation (GALE) 2008 evaluation. The dev and test sets were not released by the Linguistic Data Consortium (LDC), but compiled from the LDC released eval06, dev07, and dev08. There are two genres for each data set, namely broadcast news (BN) and broadcast conversation (BC). The ASR output is used for machine translation and information extraction. The automatically generated NE tags, confidence measures, and phone information are embedded in the automatically annotated ASR output. The names in the ASR reference are hand labeled. Table 1 shows the statistics of the data.

Table 1: Statistics of the data

Data set	#docs	#names	#names per doc	#mentions per name
dev08 BN	35	164	4.7	2.0
dev08 BC	35	103	2.9	1.5
test08 BN	34	173	5.1	2.1
test08 BC	34	86	2.5	2.2
eval08 BN	39	99	2.5	2.0
eval08 BC	34	59	1.7	1.4
Average	35	114	3.2	1.9

The number of names in BN data is almost twice as many as in BC data. Since we want to run clustering on the document level, the number of names per document should be at least greater than one. The mentions of a name also need to be more than once for that name to be possibly corrected. Ideally, larger numbers of names per document and mentions per name are better for forming clusters, but this algorithm can perform well even if these numbers are small. The average number of mentions per name is about two. However, its distribution is not uniform (Fig. 1). Due to the GALE evaluation paradigm, reference transcripts are only available for manually segmented sections (called "snippets") of a program. Most snippets are less than two minutes, consistent with the length of short news updates. The longer news commentary would potentially have more names per document and mentions per name.

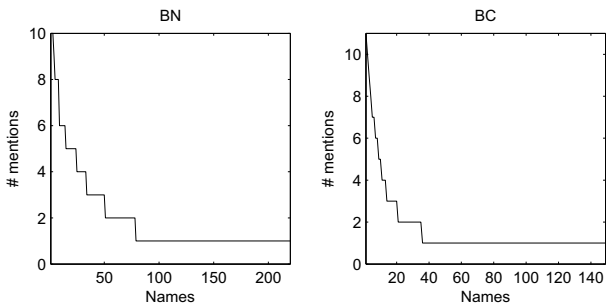

Figure 1: The distribution of the number of mentions categorized by genres, from dev08, test08 and eval08 sets.

3.2. Recognition and named entity system

Our ASR system is the multi-pass, multi-system-combined Mandarin speech recognizer described in [1, 2]. This system employs acoustic features including Mel-frequency cepstral coefficients, multi-layer perceptron, perceptual linear prediction, and smoothed pitch features. Maximum likelihood training, minimum phone error (MPE) training, and feature space MPE transform are used in acoustic model training. Cross-adaptation is applied for speaker adaptation. It also uses static and dynamic language adaptation. This system has very good overall performance. Its CER on the LDC released dev08 is 2.6% for BN, 12.1% for BC, and 7.5% on average.

To evaluate the performance of the first stage, name detection, we evaluate the NE detection on ASR output only looking at the extent of the detected names, rather than their content. The average performance of automatic NE detection on all data sets is: 98% precision, 54% recall, and 68% F-measure. It is tuned towards high precision in part to meet the needs of information extraction and machine translation. In addition, ASR errors tend to impact recall more than precision.

3.3. Experimental results

Experiments were carried out with all eight combinations of oracle and automatic components. Since the names only account for a small fraction of all the characters in the data, measuring overall CER is not appropriate. Hence, the CERs of the reference and detected names in different test sets are listed in table 2. When computing name CER, we used the union of the names that are hand labeled or automatically detected, so that the experimental results of oracle and automatic components are comparable. Therefore, the "name CER" includes a small amount of words that are actually not names in the reference.

Since we use unsupervised clustering, and the NE detection and confidence measure are all extracted from the annotated ASR output, no training is needed. However, the threshold θ that controls the determination of the number of clusters still needs to be set. It was set to a value of $\theta = 0.3$, based on the minimization of clustering error on the dev08 BN and BC sets.

3.4. Discussion

In the following we refer to different oracle experiments by their experiment IDs, which can be found in the first row of table 2. Comparing experiments 8 and 9, we see the average name CER is reduced from 22.2% to 21.3% (relative reduction 4%). The biggest name CER reduction is observed on test08 BC, from 22.9% to 20.0% (relative reduction 12%). None of the data sets have name CER worsened. From experiments 1 and 9, we can see if every component of our algorithm uses oracle information, about 23% of the ASR name error can be corrected. This is the limit of the possible name recognition improvement using this method, which is constrained by the number of the wrong name hypotheses that have correct recognitions elsewhere in the same document. These name hypotheses account for about 17% of all the names. The automatic algorithm is achieving 19% of the oracle amount.

By comparing the results within different oracle settings, we have the following observations:

- The biggest single component degradation in moving from oracle to automatic (compare 1 vs. 2, 3 and 5, with average absolute name CER change +0.7%, +2.9% and +3.8%) is for NE detection (5), as expected because of

Table 2: Name correction results (name CER)

Experiment ID		1	2	3	4	5	6	7	8	9 (no correction)
Oracle NE detection		Yes	Yes	Yes	Yes	No	No	No	No	–
Oracle name clustering		Yes	Yes	No	No	Yes	Yes	No	No	–
Oracle name hypothesis selection		Yes	No	Yes	No	Yes	No	Yes	No	–
Tuning sets	dev08 BN	12.6%	16.2%	15.8%	19.2%	17.8%	17.8%	17.8%	17.8%	18.2%
	dev08 BC	26.6%	26.6%	30.8%	30.8%	31.2%	31.2%	31.2%	31.2%	31.2%
Test sets	test08 BN	9.6%	10.6%	11.9%	12.9%	12.5%	13.5%	12.7%	13.7%	14.6%
	test08 BC	15.2%	17.1%	18.6%	18.6%	20.0%	20.0%	20.0%	20.0%	22.9%
	eval08 BN	10.4%	10.4%	12.0%	12.0%	13.4%	13.4%	13.4%	13.4%	13.4%
	eval08 BC	33.6%	33.6%	38.0%	38.0%	38.0%	38.0%	38.0%	38.0%	38.0%
test/eval average		17.2%	17.9%	20.1%	20.4%	21.0%	21.2%	21.0%	21.3%	22.2%

the low recall rates. The next largest single effect is in name clustering (3).

- There is little degradation due to name clustering when automatic name detection is used (5 vs. 7, 6 vs. 8, with name CER change +0.0% and +0.1%), since names not detected cannot lead to clustering errors.
- The automatic name hypothesis selection has a relatively small degradation compared to the oracle case when in combination with another automatic stage (3 vs. 4, 5 vs. 6, 7 vs. 8, with name CER change +0.3%, +0.2%, and +0.3%), which is likely due to the fact that when the recognition is wrong and has high confidence it is not likely to be included in the cluster.

The NE detection performance measured only in terms of extent (as in Sec. 3.2) will not be affected by name correction, as name content changes are ignored. However, the overall performance of NE recognition in speech should be evaluated based on both the location and content of detected names: A name is considered correctly recognized if and only if both its location and content match the reference. These is little impact on NE recognition for dev08 and eval08 using this metric. But for test08, both precision and recall are improved, and F-measure is improved by 4% absolute for both BN and BC.

We conjecture that the two biggest limiting factors in performance, name detection and clustering, can be attributed mainly to ASR problems. More accurate phone transcriptions, which may be achieved by improved pronunciation modeling would benefit both modules.

4. Discussions

In summary, we present a method for correcting name errors in ASR transcripts given associated phone transcripts and word confidences. The approach involves detecting names in the ASR output, clustering them at the document level, and retranscribing to enforce cluster consistency based on word confidence. Name CER has been reduced by 4% relative on average and up to 12% for the best case data set. Further experiments using oracle components have shown that name detection and clustering are the biggest factors limiting achievement of the oracle gain of 23%, but we hypothesize that improvements to name pronunciation modeling would reduce this gap.

Limiting factors in the oracle gains are the use of only 1-best recognition hypothesis and the fact that corrections are only possible when there are multiple mentions of a name. These issues can be addressed by using confusion networks to provide more name hypotheses and by augmenting the candidate name lists with information from related documents. Another possible extension would be to operate at a high recall rate and incorporate name detection posteriors into the clustering process.

5. Acknowledgments

We would like to thank Heng Ji from City University of New York for providing the human labeled name reference and automatic name tags. This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 (approved for public release, distribution unlimited). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

6. References

[1] Hwang, M.-Y., Peng, G., Wang, W., Faria, A., Heidel, A. and Ostendorf, M., "Building a highly accurate Mandarin speech recognizer," in Proceedings of ASRU'07, pp. 490-495, 2007.

[2] Lei, X., Wu, W., Wang, W., Mandal, A. and Stolcke, A., "Development of the 2008 SRI Mandarin speech-to-text system for broadcast news and conversations," in Proceedings of INTERSPEECH'09, 2009.

[3] Palmer, D. and Ostendorf, M., "Improving out-of-vocabulary name resolution," in Computer Speech & Language, vol. 19, no. 1, pp. 107-128, January 2005.

[4] Liu, Y. and Liu, F., "Unsupervised language model adaptation via topic modeling based on named entity hypotheses," in Proceedings of ICASSP'08, pp. 4921-4924, 2008.

[5] Stoyanchev, S., Tur, G. and Hakkani-Tur, D., "Name-aware speech recognition for interactive question answering," in Proceedings of ICASSP'08, pp. 5113-5116, 2008.

[6] Ji, H., Grishman, R. and Wang, W., "Phonetic name matching for cross-lingual spoken sentence retrieval," in Proceedings of IEEE-ACL SLT'08, 2008.

[7] von Luxburg, U., "A tutorial on spectral clustering," in Statistics and Computing, vol. 17, no. 4, pp. 395-416, December 2007.

[8] Ji, H., Meyers, A. and Grishman, R., "NYU's Chinese ACE 2005 EDR system description," in ACE05 PI/Evaluation Workshop. Washington, 2005. Online: http://www.cs.nyu.edu/hengji/ACE05-NYUChinese.pdf.

[9] Srinivasan, S. and Petkovic, D., "Phonetic confusion matrix based spoken document retrieval," in Proceedings of ACM SIGIR'00, pp. 81-87, 2000.

[10] Ng, A. Y., Jordan, M. I. and Weiss, Y., "On spectral clustering: analysis and an algorithm," in Advances in Neural Information Processing Systems, vol. 14, pp. 849-856, 2001.