Multi-scale Personalization for Voice Search Applications

Daniel Bolaños
Center for Spoken Language Research
University of Colorado at Boulder, USA
bolanos@cslr.colorado.edu

Geoffrey Zweig
Microsoft Research
One Microsoft Way, Redmond, WA 98052
gzweig@microsoft.com

Patrick Nguyen
Microsoft Research
One Microsoft Way, Redmond, WA 98052
panguyen@microsoft.com

Abstract

Voice Search applications provide a very convenient and direct access to a broad variety of services and information. However, due to the vast amount of information available and the open nature of the spoken queries, these applications still suffer from recognition errors. This paper explores the utilization of personalization features for the post-processing of recognition results in the form of n-best lists. Personalization is carried out from three different angles: short-term, long-term and Web-based, and a large variety of features are proposed for use in a log-linear classification framework.

Experimental results on data obtained from a commercially deployed Voice Search system show that the combination of the proposed features leads to a substantial sentence error rate reduction. In addition, it is shown that personalization features which are very different in nature can successfully complement each other.

1 Introduction

Search engines are a powerful mechanism to find specific content through the use of queries. In recent years, due to the vast amount of information available, there has been significant research on the use of recommender algorithms to select what information will be presented to the user. These systems try to predict what content a user may want based not only on the user's query but on the user's past queries, history of clicked results, and preferences. In (Teevan et al., 1996) it was observed that a significant percent of the queries made by a user in a search engine are associated to a repeated search. Recommender systems like (Das et al., 2007) and (Dou et al., 2007) take advantage of this fact to refine the search results and improve the search experience.

In this paper, we explore the use of personalization in the context of voice searches rather than web queries. Specifically, we focus on data from a multimodal cellphone-based business search application (Acero et al., 2008). In such an application, repeated queries can be a powerful tool for personalization. These can be classified into short and long-term repetitions. Short-term repetitions are typically caused by a speech recognition error, which produces an incorrect search result and makes the user repeat or reformulate the query. On the other hand, long-term repetitions, as in text-based search applications, occur when the user needs to access some information that was accessed previously, for example, the exact location of a pet clinic.

This paper proposes several different user personalization methods for increasing the recognition accuracy in Voice Search applications. The proposed personalization methods are based on extracting short-term, long-term and Web-based features from the user's history. In recent years, other user personalization methods like deriving personalized pronunciations have proven successful in the context of mobile applications (Deligne et al., 2002).

The rest of this paper is organized as follows: Section 2 describes the classification method used for rescoring the recognition hypotheses. Section 3 describes the proposed personalization methods. Section 4 describes the experiments carried out. Finally,

conclusions from this work are drawn in section 5.

2 Rescoring procedure

2.1 Log linear classification

Our work will proceed by using a log-linear classifier similar to the maximum entropy approach of (Berger and Della Pietra, 1996) to predict which word sequence W appearing on an n-best list N is most likely to be correct. This is estimated as

$$P(W|N) = \frac{\exp(\sum_i \lambda_i f_i(W, N))}{\sum_{W' \in N} \exp(\sum_i \lambda_i f_i(W', N))}. \quad (1)$$

The feature functions $f_i(W, N)$ can represent arbitrary attributes of W and N. This can be seen to be the same as a maximum entropy formulation where the class is defined as the word sequence (thus allowing potentially infinite values) but with sums restricted as a computational convenience to only those class values (word strings) appearing on the n-best list. The models were estimated with a widely available toolkit (Mahajan, 2007).

2.2 Feature extraction

Given the use of a log-linear classifier, the crux of our work lies in the specific features used. As a baseline, we take the hypothesis rank, which results in the 1-best accuracy of the decoder. Additional features were obtained from the personalization methods described in the following section.

3 Personalization methods

3.1 Short-term personalization

Short-term personalization aims at modeling the repair/repetition behavior of the user. Short-term features are a mechanism suitable for representing negative evidence: if the user repeats a utterance it normally means that the hypotheses in the previous n-best lists are not correct. For this reason, if a hypothesis is contained in a preceding n-best list, that hypothesis should be weighted negatively during the rescoring.

A straightforward method for identifying likely repetitions consists of using a fixed size time window and considering all the user queries within that window as part of the same repetition round. Once an appropriate window size has been determined, the proposed short-term features can be extracted for each hypothesis using a binary tree like the one depicted in figure 1, where feature values are in the leaves of the tree.

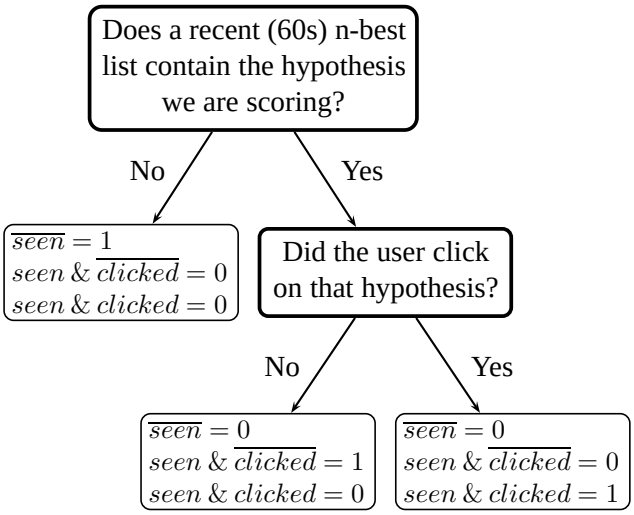

Figure 1: Short-term feature extraction (note that overlines mean "do not").

Given these features, we expect "seen and not clicked" to have a negative weight while "seen and clicked" should have a positive weight.

3.2 Long-term personalization

Long-term personalization consists of using the user history (i.e. recognition hypotheses that were confirmed by the user in the past) to predict which recognition results are more likely. The assumption here is that recognition hypotheses in the n-best list that match or "resemble" those in the user history are more likely to be correct. The following list enumerates the long-term features proposed in this work:

- User history (occurrences): number of times the hypothesis appears in the user history.
- User history (alone): 1 if the hypothesis appears in the user history and no other competing hypothesis does, otherwise 0.
- User history (most clicked): 1 if the hypothesis appears in the user history and was clicked more times than any other competing hypothesis.
- User history (most recent): 1 if the hypothesis appears in the user history and was clicked

more recently than any other competing hypothesis.

- User history (edit distance): minimum edit distance between the hypothesis and the closest query in the user history, normalized by the number of words.
- User history (words in common): maximum number of words in common between the hypothesis and each of the queries in the user history, normalized by the number of words in the hypothesis.
- User history (plural/singular): 1 if either the plural or singular version of the hypothesis appears in the user history, otherwise 0.
- Global history: 1 if the hypothesis has ever been clicked by any user, otherwise 0.
- Global history (alone): 1 if the hypothesis is the only one in the n-best that has ever been clicked by any user, otherwise 0.

Note that the last two features proposed make use of the "global history" which comprises all the queries made by any user.

3.3 LiveSearch-based features

Typically, users ask for businesses that exist, and if a business exists it probably appears in a Web document indexed by Live Search (Live Search, 2006). It is reasonable to assume that the relevance of a given business is connected to the number of times it appears in the indexed Web documents, and in this section we derive such features.

For the scoring process, an application has been built that makes automated queries to Live Search, and for each hypothesis in the n-best list obtains the number of Web documents in which it appears. Denoting by x the number of Web documents in which the hypothesis (the exact sequence of words, e.g. "tandoor indian restaurant") appears, the following features are proposed:

- Logarithm of the absolute count: $log(x)$.
- Search results rank: sort the hypotheses in the n-best list by their relative value of x and use the rank as a feature.
- Relative relevance (I): 1 if the hypothesis was not found and there is another hypothesis in the n-best list that was found more than 100 times, otherwise 0.
- Relative relevance (II): 1 if the the hypothesis appears fewer than 10 times and there is another hypothesis in the n-best list that appears more than 100 times, otherwise 0.

4 Experiments

4.1 Data

The data used for the experiments comprises 22473 orthographically transcribed business utterances extracted from a commercially deployed large vocabulary directory assistance system.

For each of the transcribed utterances two n-best lists were produced, one from the commercially deployed system and other from an enhanced decoder with a lower sentence error rate (SER). In the experiments, due to their lower oracle error rate, n-bests from the enhanced decoder were used for doing the rescoring. However, these n-bests do not correspond to the listings shown in the user's device screen (i.e. do not match the user interaction) so are not suitable for identifying repetitions. For this reason, the short term features were computed by comparing a hypothesis from the enhanced decoder with the original n-best list from the immediate past. Note that all other features were computed solely with reference to the n-bests from the enhanced decoder.

A rescoring subset was made from the original dataset using only those utterances in which the n-best lists contain the correct hypothesis (in any position) and have more than one hypothesis. For all other utterances, rescoring cannot have any effect. The size of the rescoring subset is 43.86% the size of the original dataset for a total of 9858 utterances. These utterances were chronologically partitioned into a training set containing two thirds and a test set with the rest.

4.2 Results

The baseline system for the evaluation of the proposed features consist of a ME classifier trained on only one feature, the hypothesis rank. The resulting sentence error rate (SER) of this classifier is that of the best single path, and it is 24.73%. To evaluate

the contribution of each of the features proposed in section 3, a different ME classifier was trained using that feature in addition to the baseline feature. Finally, another ME classifier was trained on all the features together.

Table 1 summarizes the Sentence Error Rate (SER) for each of the proposed features in isolation and all together respect to the baseline. "UH" stands for user history.

Features	**SER**
Hypothesis rank (baseline)	24.73%
base + repet. ($\overline{seen}$)	24.48%
base + repet. ($seen$ & $clicked$)	24.32%
base + repet. ($seen$ & $\overline{clicked}$)	24.73%
base + UH (occurrences)	**23.76%**
base + UH (alone)	23.79%
base + UH (most clicked)	23.73%
base + UH (most recent)	23.88%
base + UH (edit distance)	23.76%
base + UH (words in common)	24.60%
base + UH (plural/singular)	24.76%
base + GH	24.63%
base + GH (alone)	24.66%
base + Live Search (absolute count)	24.35%
base + Live Search (rank)	24.85%
base + Live Search (relative I)	**23.51%**
base + Live Search (relative II)	23.69%
base + all	**21.54%**

Table 1: Sentence Error Rate (SER) for each of the features in isolation and for the combination of all of them.

5 Conclusions

The proposed features reduce the SER of the baseline system by 3.19% absolute on the rescoring set, and by 1.40% absolute on the whole set of transcribed utterances.

Repetition based features are moderately useful; by incorporating them into the rescoring it is possible to reduce the SER from 24.73% to 24.32%. Although repetitions cover a large percentage of the data, it is believed that inconsistencies in the user interaction (the right listing is displayed but not confirmed by the user) prevented further improvement.

As expected, long-term personalization based features contribute to improve the classification accuracy. The UH (occurrences) feature by itself is able to reduce the SER in about a 1%.

Live Search has shown a very good potential for feature extraction. In this respect it is interesting to note that a right design of the features seems critical to take full advantage of it. The relative number of counts of one hypothesis respect to other hypotheses in the n-best list is more informative than an absolute or ranked count. A simple feature using this kind of information, like Live Search (relative I), can reduce the SER in more than 1% respect to the baseline.

Finally, it has been shown that personalization based features can complement each other very well.

References

Alex Acero, Neal Bernstein, Rob Chambers, Yun-Cheng Ju, Xiao Li, Julian Odell, Patrick Nguyen, Oliver Scholtz and Geoffrey Zweig. 2008. *Live Search for Mobile: Web Services by Voice on the Cellphone.* ICASSP 2008, March 31 2008-April 4 2008. Las Vegas, NV, USA.

Adam L. Berger; Vincent J. Della Pietra; Stephen A. Della Pietra. 1996. *A Maximum Entropy Approach to Natural Language Processing.* Computational Linguistics, 1996. 22(1): p. 39-72.

Abhinandan Das, Mayur Datar and Ashutosh Garg. 2007. *Google News Personalization: Scalable Online Collaborative Filtering.* WWW 2007 / Track: Industrial Practice and Experience May 8-12, 2007. Banff, Alberta, Canada.

Sabine Deligne, Satya Dharanipragada, Ramesh Gopinath, Benoit Maison, Peder Olsen and Harry Printz. 2002. *A robust high accuracy speech recognition system for mobile applications.* Speech and Audio Processing, IEEE Transactions on, Nov 2002, Volume: 10, Issue: 8, On page(s): 551- 561.

Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. *A large-scale evaluation and analysis of personalized search strategies.* In WWW '07: Proceedings of the 16th international conference on World Wide Web, pages 581 - 590, New York, NY, USA, 2007. ACM Press.

Live Search. "http://www.live.com,".

Milind Mahajan. 2007. *Conditional Maximum-Entropy Training Tool* http://research.microsoft.com/en-us/downloads/9f199826-49d5-48b6-ba1b-f623ecf36432/.

Jaime Teevan, Eytan Adar, Rosie Jones and Michael A. S. Potts. 2007. *Information Re-Retrieval: Repeat Queries in Yahoos Logs.* SIGIR, 2007.