Retrieving speech correctly despite the recognition errors

Mikko Kurimo and Ville Turunen

Helsinki University of Technology, FI-02015 Espoo, Finland,
Mikko.Kurimo@hut.fi,
WWW home page: http://www.cis.hut.fi/mikkok/speech/

Abstract. This work studies ways to recover from speech recognition errors in retrieving spoken documents. The methods are evaluated by Finnish news reading data using an unlimited vocabulary recognizer with language models for unsupervised morpheme-like units. Recognition errors can naturally be reduced by improving the recognizer, but the focus here is on the attempts to improve the search index more directly, namely by adding new index terms to replace those lost due to recognition errors.

1 Spoken document retrieval (SDR)

A fundamental difference between retrieving spoken and written documents is that for the speech, the index is usually based on an output of an automatic speech recognizer. This implies that the retrieval has to cope with a significant word error rate. The more spontaneous and noisy the speech is, the more errors. Typically, the error rate in different speech sections varies between 20–50 %.

Naturally, the SDR may be improved by just improving the speech recognition. The performance of our recognizer has, indeed, recently improved significantly by implementing a new decoder that efficiently operates high-order n-grams and cross-word phoneme models and new context-dependent triphones including explicit duration modeling. However, it is still possible to get even better results with less efforts by focusing on improving the index directly. In this paper we examine three advanced indexing approaches: A subword index, document expansion by pruned terms, and query expansion with text corpora.

The main novelty of this paper is the study of the ways to recover from speech recognition errors in the framework of unlimited vocabulary speech recognition based on morpheme-like language units instead of words. It continues to evaluate the SDR in the same new Finnish database as in MLMI'04 [1] reflecting the performance in an highly inflected, compounding and agglutinative language.

2 Using extra output from the speech recognizer

In information retrieval stemming is often used to associate all the inflected word forms to the same index term. In English the inflections and suffixes may not be a common source for speech recognition errors, but in highly inflected and particularly in the agglutinative languages with long words this is important.

In [1] it was already demonstrated that the addition of morpheme-based index terms to the standard word-based index, improved the retrieval performance remarkably for spoken documents. In a morpheme-based speech recognition this also comes with no additional cost, because the subword units can be directly obtained from the recognizer. Actually, using the morpheme-based output directly avoids a portion of recognition errors that result from the incorrect word segmentation. By concatenating both the morpheme and word indexes we may also recover some of the partly misrecognized words.

In addition to the morpheme description of the most likely result hypothesis, it is possible to also add other well-matching words or morphemes. In this way those misrecognized keywords or morphemes can be recovered that the recognizer found likely, but still pruned out, because the incorrect result was more likely. For this we applied a direct document expansion method that just concatenates the nearest missed morphemes to the transcription weighted by their probability. In the experiments this improved the retrieval precision in our test queries, but not as significantly as the query expansion described below.

3 Using parallel text corpora

Because some important index terms, such as foreign names, are very difficult to recognize, the best approach to retrieve the audio may be to first search parallel text collections for other terms that usually appear in the same context. This may also provide synonyms or other relevant new index terms for other query terms, which is likely to improve the retrieval precision, as well. Naturally, the final SDR performance depends on the quality and match of the chosen text corpora. The expanded terms can be also be chosen by different ways. Here we measured the relevance of the terms in the documents obtained in the first search by the frequency of the term in the retrieved documents vs. all documents, and choose the best ones for the expanded query [2].

In this work we compared the SDR performance of query expansion on two different text corpora. The other matching better to the style and category of the spoken news and the other better to the correct time block but only containing very short news summaries. We found out that the corpus matching the correct time block did better, but in both corpora the SDR improved significantly, even when using the reference transcripts instead of the recognition output. However, the final performance with the reference transcripts was about the same as with the recognized one, so we can conclude that in this evaluation the recovery from recognition errors worked very well.

References

1. Kurimo, M., Turunen, V., Ekman, I.: Speech transcription and spoken document retrieval in finnish. In: In Machine Learning for Multimodal Interaction, Revised Selected Papers of the MLMI 2004 workshop. Lecture Notes in Computer Science, vol. 3361. Springer (2005) 253–262
2. Renals, S., Abberley, D., Kirby, D., Robinson, T.: Indexing and retrieval of broadcast news. Speech Communication **32** (2000) 5–20