Spoken Information Retrieval for Turkish Broadcast News

Sıddıka Parlak
Rutgers University
Electrical and Computer Engineering Dept.
Piscataway, New Jersey 08854-8018
parlak@eden.rutgers.edu

Murat Saraçlar
Boğaziçi University
Electrical and Electronics Engineering Dept.
Bebek 34342, Istanbul, Turkey
murat.saraclar@boun.edu.tr

ABSTRACT

Speech Retrieval systems utilize automatic speech recognition (ASR) to generate textual data for indexing. However, automatic transcriptions include errors, either because of out-of-vocabulary (OOV) words or due to ASR inaccuracy. In this work, we address spoken information retrieval in Turkish, a morphologically rich language where OOV rates are high. We apply several techniques, such as using subword units and indexing alternative hypotheses, to cope with the OOV problem and ASR inaccuracy. Experiments are performed on our Turkish Broadcast News (BN) Corpus which also incorporates a spoken IR collection. Results indicate that word segmentation is quite useful but the efficiency of indexing alternative hypotheses depends on retrieval type.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; 1.2.7 [**Artificial Intelligence**]: Natural Language Processing—*speech recognition and synthesis*

General Terms

Algorithms, Languages, Performance

Keywords

Speech retrieval, Subword indexing, Spoken Term Detection, Spoken Document Retrieval

1. INTRODUCTION

Turkish is a challenging language for automatic transcription and retrieval of speech since its agglutinative structure results in high OOV rates. In this work, we examine the speech retrieval problem for Turkish focusing on two subfields: Spoken Term Detection (STD) and Spoken Document Retrieval (SDR).

The aim of STD is to locate the occurrences of a query, which requires additional methods if the query is OOV or misrecognized. Subword based indexes and word-subword hybrids have been shown to be very helpful to decrease the OOV rate [6, 7]. Indexing alternative ASR hypotheses is a successful approach to locate the misrecognized queries [7]. These alternative hypotheses can be in the form of lattices [7] or confusion networks (CN) [5].

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

Unlike STD, SDR is robust to recognition errors as reported in [4]. On the other hand, word segmentation methods are quite useful for stemming purposes [5].

IR studies on Turkish are mostly text-based. A recent work presents the first large scale text-based Turkish IR experiments on a TREC-like test collection [2].

2. DATA

A large Turkish BN Database has been collected at Boğaziçi University since March 2006 [1]. Currently, the database includes approximately 277 hours of transcribed speech.

A 74 hour portion (135 programmes) is also segmented into 2425 news stories manually and labeled with a topic. For this purpose, 27 topics are defined in short and terse forms. A short topic is generally in the structure (and length) of a sentence and a terse topic is its keyword-based counterpart. An example is presented below, in both forms:

Short Topic:
`Türkiye'de ve dünyada son zamanlarda gerçekleşmiş uçak kaçırma, hava korsanlığı vakalarını bul.`
(Find the recent skyjacking cases in Turkey and the world.)
Terse Topic:
`Uçak kaçırma hava korsanlığı` (Skyjacking cases)

3. SUBWORD-BASED UNITS

In this work, statistical and grammatical algorithms are used for word segmentation. Subwords are used as language modeling (LM) units in ASR and indexing units for retrieval.

Morphemes are grammatical units (stem and suffixes) that are extracted by a morphological parser. By combining the suffixes, we obtain 2-piece units called *grammatical stem-endings (G-SE). Morphs* are statistical subword units discovered by the unsupervised Morfessor algorithm based on the Minimum Description Length principle [3]. *Statistical stem-endings (S-SE)* are generated by grouping the second and succeeding morphs as the ending. To illustrate the difference, various parses of the word sequence "dünya kupası finalinde" are given below:

```
Word : dünya kupası finalinde
G-SE : dünya kupa +sı final +inde
Morph: dün +ya kupası final +in +de
S-SE : dün +ya kupası final +inde
```

To perform stemming, the units are slightly changed for indexing in SDR and defined as follows: *No Stemming (NS):* Each word as a whole, *Fixed Prefix (FP):* The first n characters of a word ($n = 5$) [2], *G-Stem:* First portion of G-SE, *S-Stem:* First portion of S-SE.

4. SPOKEN TERM DETECTION

Automatic transcripts (in terms of lattices) are indexed and retrieved via Weighted Finite State Transducer operations for STD. The retrieved information consists of the occurrence times of the query and the corresponding relevance scores. The detection is based on comparing the relevance scores to a threshold [7].

Performance of the STD system is evaluated with "STDEval Toolkit" in terms of the Maximum Term Weighted Value (MTWV) [6] metric. Statistics of the 3-hour test set are shown in Table 1, where WER is the word error rate, OOV shows the percentage of OOV tokens and OOV-q is the percentage of OOV queries.

The effect of subword based retrieval is investigated over all, IV (in-vocabulary) and OOV queries. As shown in Table 1, subword units provide a significant gain by locating OOV queries. If the best performer of the IV and OOV sets (words and morphs) are cascaded, MTWV increases up to 64.75.

Table 1: Statistics and STD performance of various units

	Statistics (%)			MTWV (%)		
	WER	OOV	OOV-q	all	IV	OOV
Word	26.9	1.9	27.3	56.71	**72.33**	-
Morph	26.1	0.6	7.6	60.62	67.74	**41.02**
S-SE	25.6	1.0	11.8	60.63	69.23	35.86
G-SE	25.7	0.9	8.1	**62.74**	71.88	37.52

For some STD applications, it may be sufficient to match only the stem and not the whole term. For instance, search of *Ankara* (a city name) should also return inflected forms such as *Ankara'+ya* (to Ankara) and *Ankara'+da* (in Ankara). Our stem-matching strategy increased the G-SE score from 62.74 to 65.71 and morph score from 60.62 to 64.15.

5. SPOKEN DOCUMENT RETRIEVAL

In SDR, indexing and retrieval of the news stories are performed via the traditional Vector Space Modeling technique. The system is evaluated with the "trec-eval" toolkit in terms of Binary Preference (BPref). The document collection consists of 2425 news stories. Short and terse topics (see Sec. 2) are used as queries. In addition, a third set is generated with the queries collected from users (assessors). This set gives a more realistic perspective about user requests. News stories are judged by 8 human assessors. The assessors are asked to submit their own keywords to view the news stories related to a given topic. Search is performed several times (runs) with various methods. The results are pooled and the top documents of the pool are displayed to the assessor [4].

Subword units can be used both as LM units in ASR and indexing units. Table 2 presents the SDR scores, where first row indicates the LM unit and the first column indicates the indexing unit. For example, if the LM units are G-SEs and indexing units are words (i.e. no stemming), then stems and endings in the ASR output are joined prior to indexing.

As shown in Table 2, automatic transcripts perform worse than the reference transcripts. Introducing subword based LMs does not improve the performance despite lower OOV rates and WERs. Significance tests verify our observations.

On the other hand, all stemming algorithms provide a considerable improvement. The difference between the performance of G-Stems and and S-Stems is not statistically significant. They both outperform the FP approach by creating meaningful units. Nevertheless, the difference is small enough to prefer the FP method because of its simplicity.

Table 2: BPref Scores of various transcriptions and indexing units over the user queries

	Reference	Word	G-SE	S-SE
No Stemming (NS)	38.85	37.64	38.14	37.69
Fixed-Prefix (FP)	43.15	41.73	41.63	41.37
G-Stem	43.73	41.89	41.98	-
S-Stem	43.66	42.66	-	42.94

The results above are obtained with the best recognition hypothesis. We do not notice a significant gain by indexing CNs, which are preferred to lattices because of their simplicity. The reason can be explained as follows: use of expanded hypotheses may provide the true hypotheses to be indexed and help the STD system to locate misrecognized queries. However, SDR system is already robust to small performance deviations in ASR.

6. CONCLUSIONS

This paper presents our research on STD and SDR in Turkish, as well as the collected Turkish BN Database. Indexing of grammatical and statistical subword units introduced a significant gain to STD. The stem-matching approach provided even better scores. In SDR, subwords were used separately as LM units in ASR and as indexing units in retrieval (for stemming). Although subword based ASR yielded lower WER, the overall SDR scores did not improve. On the other hand, all stemming approaches (even a very simple one: pruning to a fixed length) improved SDR scores since a common stem usually bears semantic relation. Unlike STD, indexing the alternative ASR hypotheses was not observed to be helpful for SDR.

7. ACKNOWLEDGMENTS

The authors would like to thank Ebru Arısoy, Doğan Can and Haşim Sak for the ASR system. This research is supported by TUBITAK (Project code: 105E102) and Boğaziçi University Research Fund (Project code: 05HA202).

8. REFERENCES

[1] E. Arısoy, D. Can, S. Parlak, H. Sak, and M. Saraçlar. Turkish broadcast news transcription and retrieval. *IEEE Transactions on Speech and Audio Processing*, June 2009.

[2] F. Can, S. Koçberber, E. Balçık, C. Kaynak, and H. C. Öcalan. Information retrieval on Turkish texts. *JASIST*, 59(3):407–421, February 2008.

[3] M. Creutz and K. Lagus. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Publications in Computer and Information Science Report A81, Helsinki University of Technology, March 2005.

[4] J. Garofolo, G. Auzanne, and E. Voorhees. The TREC spoken document retrieval track: A success story. In *Proc. TREC 8*, pages 16–19, 2000.

[5] M. Kurimo and V. Turunen. Indexing confusion networks for morph-based spoken document retrieval. In *Proc. SIGIR*, pages 631–638, July 2007.

[6] NIST. (STD) 2006 evaluation plan http://www.nist.gov/speech/tests/std/. 2006.

[7] S. Parlak and M. Saraçlar. Spoken term detection for Turkish broadcast news. In *Proc. ICASSP*, pages 5244–5247, April 2008.