Morphological Segmentation for Keyword Spotting

Karthik Narasimhan[1], **Damianos Karakos**[2], **Richard Schwartz**[2], **Stavros Tsakalidis**[2], **Regina Barzilay**[1]
[1]Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology
[2]Raytheon BBN Technologies
{karthikn, regina}@csail.mit.edu
{dkarakos, schwartz, stavros}@bbn.com

Abstract

We explore the impact of morphological segmentation on keyword spotting (KWS). Despite potential benefits, state-of-the-art KWS systems do not use morphological information. In this paper, we augment a state-of-the-art KWS system with sub-word units derived from supervised and unsupervised morphological segmentations, and compare with phonetic and syllabic segmentations. Our experiments demonstrate that morphemes improve overall performance of KWS systems. Syllabic units, however, rival the performance of morphological units when used in KWS. By combining morphological, phonetic and syllabic segmentations, we demonstrate substantial performance gains.

1 Introduction

Morphological analysis plays an increasingly important role in many language processing applications. Recent research has demonstrated that adding information about word structure increases the quality of translation systems and alleviates sparsity in language modeling (Chahuneau et al., 2013b; Habash, 2008; Kirchhoff et al., 2006; Stallard et al., 2012).

In this paper, we study the impact of morphological analysis on the keyword spotting (KWS) task. The aim of KWS is to find instances of a given keyword in a corpus of speech data. The task is particularly challenging for morphologically rich languages as many target keywords are unseen in the training data. For instance, in the Turkish dataset (Babel, 2013) we use, from the 2013 IARPA Babel evaluations, 36.06% of the test words are unseen in the training data. However, 81.44% of these unseen words have a morphological variant in the training data. Similar patterns are observed in other languages used in the Babel evaluations. This observation strongly supports the use of morphological analysis to handle out-of-vocabulary (OOV) words in KWS systems.

Despite this potential promise, state-of-the-art KWS systems do not commonly use morphological information. This surprising fact can be due to multiple reasons, ranging from the accuracy of existing morphological analyzers to the challenge of integrating morphological information into existing KWS architectures. While using morphemes is likely to increase coverage, it makes recognition harder due to the inherent ambiguity in the recognition of smaller units. Moreover, it is not clear a priori that morphemes, which are based on the semantics of written language, are appropriate segmentation units for a speech-based application.

We investigate the above hypotheses in the context of a state-of-the-art KWS architecture (Karakos et al., 2013). We augment word lattices with smaller units obtained via segmentation of words, and use these modified lattices for keyword spotting. We consider multiple segmentation algorithms, ranging from near-perfect supervised segmentations to random segmentations, along with unsupervised segmentations and purely phonetic and syllabic segmentations. Our experiments show how sub-word units can be used effectively to improve the performance of KWS systems. Further, we study the extent of impact of the subwords, and the manner in which they can be used in KWS systems.

2 Related Work

Prior research on applications of morphological analyzers has focused on machine translation, language modeling and speech recognition (Habash, 2008; Chahuneau et al., 2013a; Kirchhoff et al., 2006). Morphological analysis enables us to link together multiple inflections of the same root, thereby alleviating word sparsity common in mor-

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 880–885,
October 25-29, 2014, Doha, Qatar.

phologically rich languages. This results in improved language model perplexity, better word alignments and higher BLEU scores.

Recent work has demonstrated that even morphological analyzers that use little or no supervision can help improve performance in language modeling and machine translation (Chahuneau et al., 2013b; Stallard et al., 2012). It has also been shown that segmentation lattices improve the quality of machine translation systems (Dyer, 2009).

In this work, we leverage morphological segmentation to reduce OOV rates in KWS. We investigate segmentations produced by a range of models, including acoustic sub-word units. We incorporate these subword units into a lattice framework within the KWS system. We also demonstrate the value of using alternative segmentations instead of or in combination with morphemes. In addition to improving the performance of KWS systems, this finding may also benefit other applications that currently use morphological segmentation for OOV reduction.

3 Segmentation Methods

Supervised Morphological Segmentation Due to the unavailability of gold morphological segmentations for our corpus (Babel, 2013), we use a resource-rich supervised system as a proxy. As training data for this system, we use the MorphoChallenge 2010 corpus[1] which consists of 1760 gold segmentations for Turkish.

We consider two supervised frameworks, both made up of two stages. In the first stage, common to both systems, we use a FST-based morphological parser (Çöltekin, 2010) that generates a set of candidate segmentations, leveraging a large database of Turkish roots and affixes. This stage tends to overgenerate, segmenting each word in eight different ways on average. In the next stage, we filter the resulting segmentations using one of two supervised filters (described below) trained on the MorphoChallenge corpus.

In the first approach, we use a binary log-linear classifier to accept/reject each segmentation hypothesis. For each word, this classifier may accept multiple segmentations, or rule out all the alternatives. In the second approach, to control the number of segmentations per word, we train a log-linear ranker that orders the segmentations for a word in decreasing order of likelihood. In our training corpus, each word has on average 2.5 gold segmentations. Hence, we choose the top two segmentations per word from the output of the ranker to use in our KWS system. In both filters, we use several features like morpheme unigrams, bigrams, lengths, number of morphemes, and phone sequences corresponding to the morphemes.

Feature	Example
morpheme unigrams	tak, acak
morpheme bigram	⟨tak, acak⟩
phonetic seq. unigrams	t.a.k., 1v.dZ.a.k.
phonetic seq. bigram	⟨t.a.k., 1v.dZ.a.k.⟩
number of morphemes	2
morpheme lengths	3, 4

Table 1: Example of features used in the supervised filters for the segmentation *tak-acak*. Each phone is followed by a dot for clarity.

In our supervised systems, we can encode features that go beyond individual boundaries, like the total number of morphemes in the segmentation. This global view distinguishes our classifier/ranker from traditional approaches that model segmentation as a sequence tagging task (Ruokolainen et al., 2013; Kudo et al., 2004; Kruengkrai et al., 2006). Another departure of our approach is the use of phonetic information, in the form of phonetic sequences corresponding to the morpheme unigrams and bigrams. The hypothesis is that syllabic boundaries are correlated with morpheme boundaries to some extent. The phonetic sequences for words are obtained using a publicly available Text-to-Phone (T2P) system (Lenzo, 1998).

Unsupervised Morphological Segmentation We employ a widely-used unsupervised system Morfessor (Creutz and Lagus, 2005) which achieves state-of-the-art unsupervised performance in the MorphoChallenge evaluation. Morfessor uses probabilistic generative models with sparse priors which are motivated by the Minimum Description Length (MDL) principle. The system derives segmentations from raw data, without reliance on extra linguistic sources. It outputs a single segmentation per word.

Random Segmentation As a baseline, we include sub-word units from random segmentations, where we mark a segmentation boundary at each character position in a word with a fixed probability p. For comparison purposes, we consider two

[1] http://research.ics.aalto.fi/events/morphochallenge2010/

Sub-word units	Example
Morphemes	tak - acak
Random	t - aka - c - a - k
Phones	t - a - k - 1v - dZ - a - k
Syllables	ta - k1v - dZak

Table 2: Segmentations of the word *takacak* into different types of sub-word units.

types of random segmentations that match the supervised morphological segmentations in terms of the number of uniques morphemes and the average morpheme length, respectively. These segmentations are obtained by adjusting the segmentation probability p appropriately.

Phones and Syllables In addition to letter-based segmentation, we also consider other subword units that stem from word acoustics. In particular, we consider segmentation using phones and syllables, which are available for the Babel data we work with.

Table 2 shows examples of different segmentations for the Turkish word *takacak*.

4 Keyword Spotting

The keyword spotting system used in this work follows, to a large extent, the pipeline of (Bulyko et al., 2012). Using standard speech recognition machinery, the system produces a detailed lattice of word hypotheses. The resulting lattice is used to extract keyword hits with nominal posterior probability scores.

We modify this basic architecture in two ways. First, we use subwords instead of whole-words in the decoding lexicon. Second, we represent keywords using all possible paths in a lattice of subwords. For each sequence of matching arcs in the lattice, the posteriors of these arcs are multiplied together to form the score of detection (hit). A post-processing step adds up (or takes the max of) the scores of all hits of each keyword which have significant overlap in time. Finally, the hit lists are processed by the score normalization and combination method described in (Karakos et al., 2013).

We use whole-word extraction for words in vocabulary, but rely on subword models for OOV words. Since we combine the hits separately for IV and OOV keywords, using subwords can only improve the performance of the overall system.

Language	Dev Set	Eval Set
Turkish	403	226
Assamese	158	563
Bengali	176	629
Haitian	107	319
Lao	110	194
Tamil	238	700
Zulu	323	1251

Table 3: Number of OOV keywords in the different Dev and Eval sets.

5 Experimental Setup

Data The segmentation algorithms described in Section 3 are tested using the setup of the KWS system described in Section 4. Our experiments are conducted using the IARPA Babel Program language collections for Turkish, Assamese, Bengali, Haitian, Lao, Tamil and Zulu (Babel, 2013)[2]. The dataset contains audio corpora and a set of keywords. The training corpus for KWS consists of 10 hours of speech, while the development and test sets have durations of 10 and 5 hours, respectively. We evaluate KWS performance over the OOV keywords in the data, which are unseen in the training set, but appear in the development/test set. Table 3 contains statistics on the number of OOV keywords in the data for each language.

In our experiments, we consider the pre-indexed condition, where the keywords are known only after the decoding of the speech has taken place.

Evaluation Measures We consider two different evaluation metrics. To evaluate the accuracy of the different segmentations, we compare them against gold segmentations from the MorphoChallenge data for Turkish. This set consists of 1760 words, which are manually segmented. We use a measure of word accuracy (**WordAcc**), which captures the accuracy of all segmentation decisions within the word. If one of the segmentation boundaries is wrong in a proposed segmentation, then that segmentation does not contribute towards the WordAcc score. We use 10-fold cross-validation for the supervised segmentations, while we use the entire set for unsupervised and acoustic cases.

We evaluate the performance of our KWS system using a widely used metric in KWS, the Ac-

[2]We perform the experiments with supervised segmentation only on Turkish, due to the lack of gold morphological data for the other languages.

tual Term Weighted Value (ATWV) measure, as described in (Fiscus et al., 2007). This measure uses a combination of penalties for misses and false positives to score the system. The maximum score achievable is 1.0, if there are no misses and false positives, while the score can be lower than 0.0 if there are a lot of misses or false positives.

6 Results

Table 4 summarizes the performance of all considered segmentation systems in the KWS task on Turkish. The quality of the segmentations compared to the gold standard is also shown. Table 5 shows the OOV ATWV performance on the six other languages, used in the second year of the IARPA Babel project. We summarize below our conclusions based on these results.

Using sub-word units improves overall KWS performance If we use a word-based KWS system, the ATWV score will be 0.0 since the OOV keywords are not present in the lexicon. Enriching our KWS system with sub-word segments yields performance gains for all the segmentation methods, including random segmentations. However, the observed gain exhibits significant variance across the segmentation methods. For instance, the gap between the performance of the KWS system using the best supervised classifier-based segmenter (*CP*) and that using the unsupervised segmenter (*U*) is 0.059, which corresponds to a 43.7% in relative gain. Table 4 also shows that while methods with shorter sub-units (*U*, *P*) yield lower OOV rate, they do not necessarily fare better in the KWS evaluation.

Syllabic units rival the performance of morphological units A surprising discovery from our experiments is the good performance of the syllabic segmentation-based KWS system (*S*). It outperforms all the alternative segmentations on the test set, and ranks second on the development set behind the *CP* system. These units are particularly attractive as they can easily be computed from acoustic input and do not require any prior linguistic knowledge. We hypothesize that the granularity of this segmentation is crucial to its success. For instance, a finer-grained phone-based segmentation (*P*) performs substantially worse than other segmentation algorithms as the derived sub-units are shorter and hence, harder to recognize.

Improving morphological accuracy beyond a certain level does not translate into improved KWS performance We observe that the segmentation accuracy and KWS performance are not positively correlated. Clearly, bad segmentations translate into poor ATWV scores, as in the case of random and unsupervised segmentations. However, gains on segmentation accuracy do not always result in better KWS performance. For instance, the ranker systems (*RP, RNP*) have better accuracies on MC2010, while the classifier systems (*CP, CNP*) perform better on the KWS task. This discrepancy in performance suggests that further gains can be obtained by optimizing segmentations directly with respect to KWS metrics.

Adding phonetic information improves morphological segmentation For all the morphological systems, adding phonetic information results in consistent performance gains. For instance, it increases segmentation accuracy by 4% when added to the classifier (*CNP* and *CP* in table 4). The phonetic information used in our experiments is computed automatically using a T2P system (Lenzo, 1998), and can be easily obtained for a range of languages. This finding sheds new light on the relation between phonetic and morphological systems, and can be beneficial for morphological analyzers developed for other applications.

Combining morphological, phonetic and syllabic segmentations gives better results than either in isolation As table 4 shows, the best KWS results are achieved when syllabic and morphemic systems are combined. The best combination system (*CP+P+S*) outperforms the best individual system (*S*) by 5.5%. This result suggests that morphemic, phonemic and syllabic segmentations encode complementary information which benefits KWS systems in handling OOV keywords.

Morphological segmentation helps KWS across different languages Table 5 demonstrates that we can obtain gains in KWS performance across different languages using unsupervised segmentation. The improvement is significant in 3 of the 6 languages - as high as 3.2% for Assamese and Bengali, and 2.7% for Tamil (absolute percentages). As such, the results of Table 2 cannot be directly compared to those of Table 1 since the system architecture is slightly different[3]. How-

[3]The keyword spotting pipeline is based on the one used by the Babelon team in the 2014 NIST evaluation (Tsakalidis, 2014). The pipeline was much more involved than the one described for Turkish; multiple search methods (with/without fuzzy search) and data structures (lattices, confusion networks and generalized versions of these) were all used in combination (Karakos and Schwartz, 2014). The recognition

Method	Unique units	Avg. unit length	Reduction in OOV (abs)	WordAcc	Dev ATWV	Test ATWV
Phone-based (P)	51	1	36.06%	0.06%	0.099	0.164
Syllable-based (S)	2.1k	3.62	23.91%	10.29%	0.127	0.201
Classifier w/ phone info (CP)	18.5k	6.39	18.20%	80.41%	0.146	0.194
Classifier w/o phone info (CNP)	19k	6.42	21.50%	75.66%	0.133	0.181
Ranker w/ phone info (RP)	10k	5.62	16.86%	**86.03%**	0.104	0.153
Ranker w/o phone info (RNP)	10k	5.71	16.44%	84.19%	0.109	0.159
Unsupervised (U)	2.4k	5.44	22.45%	39.57%	0.080	0.135
RANDLen-Classifier	11.7k	6.39	0.73%	5.11%	0.061	0.086
RANDNum-Classifier	18.2k	3.03	8.56%	3.69%	0.111	0.154
RANDLen-Ranker	11.6k	5.62	1.94%	5.79%	0.072	0.136
RANDNum-Ranker	11.7k	6.13	1.15%	5.34%	0.081	0.116
CP + P	-	-	-	-	0.190	0.246
RP + P	-	-	-	-	0.150	0.210
CP + P + S	-	-	-	-	**0.208**	**0.257**
RP + P + S	-	-	-	-	0.186	0.249
Word-based for IV words	-	-	-	-	0.385	0.400

Table 4: Segmentation Statistics and ATWV scores on Babel Turkish data along with WordAcc on MorphoChallenge 2010 data. All rows except the last are for OOV words. Absolute reduction is from an initial OOV of 36.06%. Higher ATWV scores are better. Best system scores are shown in bold.

	Assamese		Bengali		Haitian		Lao		Tamil		Zulu	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
P + S	0.213	0.230	0.277	0.296	0.371	0.342	0.228	0.139	0.349	0.267	0.279	0.215
P + S + U	0.214	**0.263**	**0.294**	**0.328**	**0.393**	0.342	0.237	0.146	**0.395**	**0.284**	0.275	0.218

Table 5: ATWV scores for languages used in the second year of the IARPA Babel project, using two KWS systems: Phone + Syllable (P+S) and Phone + Syllable + Unsupervised Morphemes (P+S+U). Bold numbers show significant performance gains obtained by adding morphemes to the system.

ever, they are indicative of the large gains (1.5%, on average, over the six languages) that can be obtained through unsupervised morphology, on top of a very good combined phonetic/syllabic system.

7 Conclusion

We explore the extent of impact of morphological segmentation on keyword spotting (KWS). To investigate this issue, we augmented a KWS system with sub-word units derived by multiple segmentation algorithms. Our experiments demonstrate that morphemes improve the overall performance of KWS systems. Syllabic units, however, rival the performance of morphemes in the KWS task. Furthermore, we demonstrate that substantial performance gains in KWS performance are obtained by combining morphological, phonetic and syllabic segmentations. Finally, we also show that adding phonetic information improves the quality of morphological segmentation.

was done with audio features supplied by BUT (Karafiát et al., 2014), which were improved versions of those used for Turkish.

Acknowledgements

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government. We thank the MIT NLP group and the EMNLP reviewers for their comments and suggestions.

References

IARPA Babel. 2013. Language collection releases; Turkish: IARPA-babel105b-v0.4, Assamese: IARPA-babel102b-v0.5a, Bengali: IARPA-babel103b-0.4b, Haitian Creole: IARPA-babel201b-v0.2b, Lao: IARPA-babel203b-v3.1a, Tamil: IARPA-babel204b-v1.1b, Zulu: IARPA-babel206b-v0.1e.

Ivan Bulyko, Owen Kimball, Man-Hung Siu, José Herrero, and Dan Blum. 2012. Detection of unseen words in conversational Mandarin. In *Proc. of ICASSP*, Kyoto, Japan, Mar.

Victor Chahuneau, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013a. Translating into morphologically rich languages with synthetic phrases. In *EMNLP*, pages 1677–1687. ACL.

Victor Chahuneau, Noah A. Smith, and Chris Dyer. 2013b. Knowledge-rich morphological priors for bayesian language models. In *HLT-NAACL*, pages 1206–1215. The Association for Computational Linguistics.

Çağrı Çöltekin. 2010. A freely available morphological analyzer for Turkish. In *Proceedings of the 7th International conference on Language Resources and Evaluation (LREC2010)*, pages 820–827.

Mathias Creutz and Krista Lagus. 2005. Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR)*, pages 106–113.

Chris Dyer. 2009. Using a maximum entropy model to build segmentation lattices for MT. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 406–414, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jonathan G. Fiscus, Jerome Ajot, John S. Garofolo, and George Doddington. 2007. Results of the 2006 spoken term detection evaluation. In *Workshop on Searching Spontaneous Conversational Speech*.

Nizar Habash. 2008. Four techniques for online handling of out-of-vocabulary words in arabic-english statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 57–60, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martin Karafiát, František Grézl, Mirko Hannemann, Karel Veselý, Igor Szoke, and Jan "Honza" Černocký. 2014. BUT 2014 Babel system: Analysis of adaptation in NN based systems. In *Proceedings of Interspeech 2014*, Singapore, September. IEEE.

Damianos Karakos and Richard Schwartz. 2014. Subword modeling. In *IARPA Babel PI Meeting*, July.

Damianos Karakos, Richard Schwartz, Stavros Tsakalidis, Le Zhang, Shivesh Ranjan, Tim Ng, Roger Hsiao, Guruprasad Saikumar, Ivan Bulyko, Long Nguyen, John Makhoul, Frantisek Grezl, Mirko Hannemann, Martin Karafiat, Igor Szoke, Karel Vesely, Lori Lamel, and Viet-Bac Le. 2013. Score normalization and system combination for improved keyword spotting. In *Proc. ASRU 2013*, Olomouc, Czech Republic.

Katrin Kirchhoff, Dimitra Vergyri, Jeff Bilmes, Kevin Duh, and Andreas Stolcke. 2006. Morphology-based language modeling for conversational arabic speech recognition. *Computer Speech and Language*, 20(4):589–608.

Canasai Kruengkrai, Virach Sornlertlamvanich, and Hitoshi Isahara. 2006. A conditional random field framework for Thai morphological analysis. In *LREC*.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *In Proc. of EMNLP*, pages 230–237.

Kevin Lenzo. 1998. Text-to-phoneme converter builder. `http://www.cs.cmu.edu/afs/cs.cmu.edu/user/lenzo/html/areas/t2p/`. Accessed: 2014-03-11.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 29–37, Sofia, Bulgaria, August. Association for Computational Linguistics.

David Stallard, Jacob Devlin, Michael Kayser, Yoong Keok Lee, and Regina Barzilay. 2012. Unsupervised morphology rivals supervised morphology for Arabic MT. In *ACL (2)*, pages 322–327. The Association for Computer Linguistics.

Stavros Tsakalidis. 2014. The Babelon OpenKWS14 systems. In *IARPA Babel PI Meeting*, July.