Integrating Morphology into Automatic Speech Recognition

Haşim Sak #1, Murat Saraçlar *2, Tunga Güngör #3

# *Department of Computer Engineering, Boğaziçi University*
TR-34342, Bebek, İstanbul, Turkey
[1] hasim.sak@boun.edu.tr
[3] gungort@boun.edu.tr

* *Department of Electrical & Electronics Engineering, Boğaziçi University*
TR-34342, Bebek, İstanbul, Turkey
[2] murat.saraclar@boun.edu.tr

***Abstract*—This paper proposes a novel approach to integrate the morphology of a language into an automatic speech recognition (ASR) system. The high out-of-vocabulary (OOV) word rates have been a major challenge for ASR in morphologically productive languages. The standard approach to this problem has been to shift from words to sub-word units. We propose to integrate the morphology as any other knowledge source - such as lexicon, and language model - directly into the search network. The morphological parser for a language, implemented as a finite-state lexical transducer can be considered as a computational lexicon. The computational lexicon presents a dynamic vocabulary in contrast to a static vocabulary generally used for ASR in morphologically poor languages. We compose the transducer for this computational lexicon with a statistical language model over lexical morphemes to obtain a morphology-integrated search network. We give experimental results for Turkish broadcast news transcription, and show that it outperforms the 50K and 100K vocabulary word models while the 200K vocabulary word model is slightly better.**

I. Introduction

The morphologically productive languages such as Turkish, Finnish, and Czech present some challenges in automatic speech recognition (ASR) systems. The out-of-vocabulary (OOV) rates for a fixed vocabulary size are significantly higher in these languages. The higher OOV rates lead to higher word error rates (WERs). Having a large number of words also contributes to high perplexity numbers for standard n-gram language models. Turkish, being an agglutinative language with a highly productive inflectional and derivational morphology is especially prone to these problems.

We can reduce the OOV rate by increasing the vocabulary size. However, this increases the computational and memory requirements of the system. It may also lead to data sparsity resulting in non-robust language models. Therefore, to overcome the OOV problem, using sub-word units for language modeling has been a common approach [1], [2]. The sub-word units can be morphological units, or some grouping of them such as stem and ending (grouping of suffixes). They can also be obtained by splitting words using statistical methods. Sub-word language models alleviate the OOV problem, however the speech decoder can generate ungrammatical sub-word sequences and post-processing of the sub-word lattices are required to correct the errors and increase the accuracy [3], [4].

Using a fixed vocabulary in ASR systems for morphologically complex languages is not the optimal approach. In such languages, grammatical features and functions associated with the syntactic structure of a sentence in morphologically poor languages, are often represented in the morphological structure of a word in addition to the syntactic structure. Therefore, morphological parsing of a word may reveal valuable information for language modeling in its constituent morphemes annotated with morphosnytactic and morphosemantic features.

The weighted finite-state transducer (WFST) provides a unified framework for representing different knowledge sources in ASR systems, e.g., hidden Markov models (HMMs), context-dependent dependency networks, pronunciation lexicons, and n-gram language models [5]. The WFST also offers finite-state operations such as composition, determinization and minimization to combine all these knowledge sources into an optimized all-in-one search network. In this framework, morphology can also be considered as another knowledge source to be integrated into the search network.

In this paper, we propose a novel approach to integrate the morphology of a language into the WFST framework of ASR systems. The morphological information in a word can be extracted using a morphological parser. The finite-state transducers present enough expressive power for representing the morphology for many languages. The lexical transducer of the parser can be considered as a computational lexicon. This dynamic lexicon can be integrated into the WFST framework by composing this lexical transducer with a language model estimated over lexical morphemes to obtain a morphology-integrated search network.

The rest of the paper is organized as follows: In section II, we describe the language resources that we built for Turkish. In section III, we present the methodology to integrate the morphology into the search network. In section IV, we give experimental results for a Turkish broadcast news transcription task. In section V, we conclude with a discussion of the results.

II. Turkish Language Resources

A. Text Corpus

We compiled a text corpus from web for estimating the parameters of statistical language models [6]. The text corpus (BOUN NewsCor) contains about 200 million-words collected by crawling three major news portals in Turkish. The corpus has been cleaned automatically and about 96.7% of the tokens can be parsed using the morphological parser that we built.

In this paper, we also used a 1.3 million-words text corpus obtained from the transcriptions of the Turkish Broadcast News speech database [2].

B. Morphological Parser

We built a morphological parser using the two-level morphology formalism of Koskenniemi [7]. In two-level morphology, the finite-state transducer of the morphological parser is obtained as the composition of the morphophonemics transducer, which encodes the phonological rules such as for vowel harmony phenomena, and the morphotactics transducer, which encodes the morphosyntax of the language. The two-level phonological rules and the morphotactics were adapted from the PC-KIMMO implementation of Oflazer [8]. The rules were compiled using the *twolc* rule compiler [9]. A new root lexicon of 55,278 words based on the Turkish Language Institution dictionary[1] was compiled. For finite-state operations, we used the OpenFST weighted finite-state transducer library [10]. The parser can analyze about 8700 words per second on a 2.33 GHz Intel Xeon processor.

The (inverse of) morphological parser can generate infinite number of words, due to iteration of some suffixes such as causation and noun-verb-noun cycles in morphotactics of the language. However, the words having more than 6 morphemes are rarely used in practice. For instance, the BOUN NewsCor text corpus contains about 1 million unique morphological analyses for about the 1.5 million unique words - disambiguated with a morphological disambiguator. Figure 1 shows the histogram for the morpheme counts in these analyses. There are only 21 analyses with the morpheme count 9. One of these is the analyses for the word *ölümsüzleştirilebileceğini*:

ölüm[Noun]-*sHz*[Adj+Without]-*lAş*[Verb+Become]
-*DHr[Verb+Caus]*-*Hl*[Verb+Pass]-*YAbil[Verb+Able]*
-*YAcAk*[Noun+FutPart]+*SH*[P3sg]+*NH*[Acc]

This word can be translated as *"... that s/he can be immortalized"*.

For this work, we modified the morphotactics of the parser to remove the morphological features which are not associated with a lexical morpheme, thus enabling us to build a language model over lexical morphemes. Figure 2 shows the part of the lexical transducer of this modified morphological parser for the ambiguous word *haberleri*. The morphological features are identical to the ones in [11].

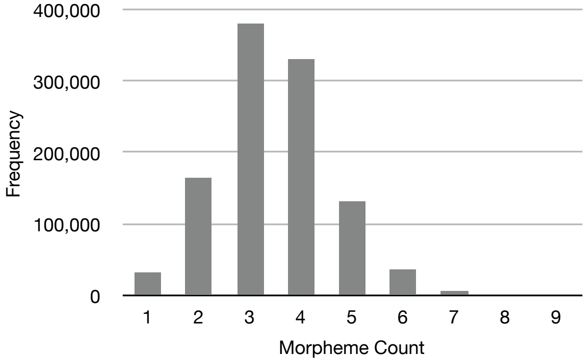

Fig. 1. The histogram for the morpheme counts

C. Morphological Disambiguator

The morphological parser may output more than one possible analysis for a word due to ambiguity. For example, the parser returns four analyses for the ambiguous word *haberleri* as shown below.

haber[Noun]+*lAr*[A3pl]+*SH*[P3sg]+[Nom] (his/her news)
haber[Noun]+*lAr*[A3pl]+[Pnon]+*YH*[Acc] (the news)
haber[Noun]+*lAr*[A3pl]+*SH*[P3pl]+[Nom] (their news)
haber[Noun]+[A3sg]+*lArH*[P3pl]+[Nom] (their (singular) news)

We need to resolve this ambiguity to estimate statistical morpheme-based models. For this purpose, we used our averaged perceptron-based morphological disambiguator [6]. The disambiguation system achieves about 97.05% disambiguation accuracy on the test set.

III. Integrating Morphology

In the WFST framework, the speech recognition problem is treated as a transduction from input speech signal to a word sequence. The various knowledge sources are represented as WFSTs. A typical set of knowledge sources consists of a hidden Markov model H mapping HMM state ID sequences to context-dependent phones, a context-dependency network C transducing context-dependent phones to context-independent phones, a lexicon L mapping context-independent phone sequences to words, and a language model G assigning probabilities to word sequences. The composition of these models $H \circ C \circ L \circ G$ results in all-in-one search network that directly maps HMM state ID sequences to weighted word sequences.

The morphology as another knowledge source can be represented as a WFST and can be integrated into the WFST framework of an ASR system. The lexical transducer of the morphological parser maps the letter sequences to lexical morphemes annotated with morphological features. The lexical transducer can be considered as a computational dynamic lexicon in ASR in contrast to a static lexicon. The computational lexicon has some advantages over a fixed-size word lexicon. It can generate many more words using a relatively smaller number of root words in its lexicon. So that it achieves lower

[1] http://www.tdk.gov.tr

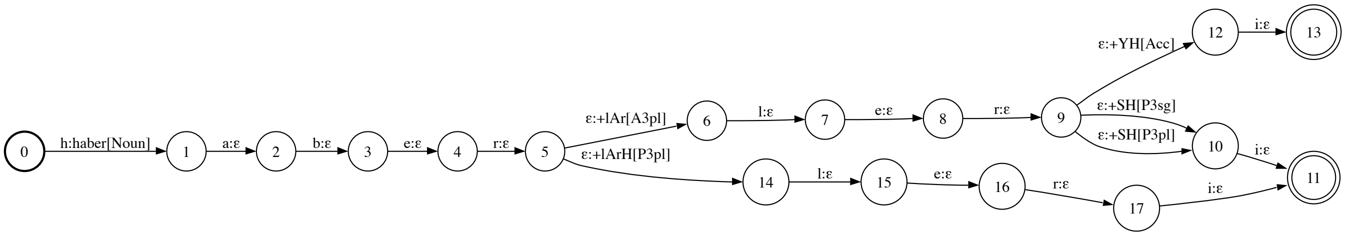

Fig. 2. Part of the lexical transducer of the parser for the word *haberleri*

OOV rates. For example, the BOUN NewsCor corpus has an average number of 30 different words derived from each root word in the lexicon. Different than the static lexicon, even if we have never seen a specific word in the training corpus, the speech decoder has the chance to recognize that word. Another benefit of the computational lexicon is that it outputs the morphological analysis of the word generated. We can exploit this morphological information in a language model.

Since Turkish is a phonetic language, we use the Turkish letters as our phone set in Turkish ASR. Note that one can also build a finite-state transducer based pronunciation lexicon [11]. In the WFST framework, the lexical transducer of the morphological parser can be considered as a computational lexicon M replacing the static lexicon L. Since the M outputs lexical morphemes, the language model G should be estimated over these lexical units. Then with the morphology integrated, the search network can be built as $H \circ C \circ M \circ G_{morpheme}$.

The WFST offers finite-state operations such as *composition*, *determinization* and *minimization* to combine all the knowledge sources used in speech recognition and optimize into a single compact search network [12]. This approach works well for certain types of transducers, but presents some problems related to the applicability of *determinization* and *weight-pushing* with more general transducers [13]. In this respect, Turkish morphology presents a problem, since the number of ambiguities is infinite and the cycle-ambiguous finite-state transducer of the morphological parser is not determinizable. As can be seen in Figure 1, the number of morphemes used in practice is limited. Therefore, we can limit the lexical transducer of the parser to generate words only up to a fixed maximal number of morphemes by composing it with a length-limited lexical morpheme transducer. Since all the acyclic weighted transducers are determinizable, we expect to be able to build the optimized search network as in [13]. But the number of ambiguities is still very large preventing the application of optimization algorithms due to computational unfeasibility.

However, we can apply local determinization algorithm for locally optimizing the search network using the *grmlocaldeterminize* utility from *AT&T Grammar Library* [14]. The experimental results show that this approach works well.

IV. Experiments

We evaluated the effectiveness of the proposed approach for morphology integration on a broadcast news transcription task. The acoustic model uses hidden Markov models (HMMs) trained on 194 hours of broadcast news speech data [2]. In the acoustic model, there are 10843 triphone HMM states and 11 Gaussians per state with the exception of the 23 Gaussians for the silence HMM. The test set contains 3.1 hours of speech data (2,410 utterances).

We built three baseline word models with static vocabularies of sizes 50K, 100K and 200K. The n-gram order of the baseline word language models was chosen to be 3. The search networks for the baseline models are constructed by composing and optimizing $L \circ G$ with *dmake* utility from *AT&T DCD Library* [2]. The context-dependency and the hidden Markov models are composed on-the-fly in the decoder. The baseline 3-gram language models were estimated by linearly interpolating two language models trained over the BOUN NewsCor corpus and the corpus from the transcriptions of broadcast news training speech data to reduce the effect of out-of-domain data using the SRILM toolkit [15].

Both text corpus were morphologically parsed and disambiguated to get the lexical morphemes and to build a 5-gram lexical morpheme language model which was composed with the lexical transducer of the morphological parser $M \circ G_{morpheme}$. The search network was optimized using the local determinization.

Table I shows the OOV rates and the perplexity values on the test set for the baseline models and the morphology model. While the morphology model has a OOV rate between 100K and 200K word models, the perplexity of the model is higher than the 200K word model. This shows that the probability estimates of the 5-gram morpheme-based model is not as good as the 3-gram word model. Table II shows that the search network sizes of all the models are comparable due to entropy-based pruning of the language models with the same threshold using the SRILM toolkit [15].

TABLE I
Language Model OOV Rates and Perplexities

Model	Vocabulary Size	OOV Rate (%)	Perplexity
Word	50K	7.50	280
Word	100K	4.06	357
Word	200K	2.01	427
Morpheme	55K	2.97	531

[2] http://www.research.att.com/˜fsmtools/dcd/

TABLE II
THE SIZE OF WFSTS

Model	Vocabulary Size	# of states	# of arcs
Word	50K	7,137,968	15,145,185
Word	100K	7,881,797	16,240,845
Word	200K	8,652,109	17,325,753
Morpheme	55K	7,958,080	18,922,956

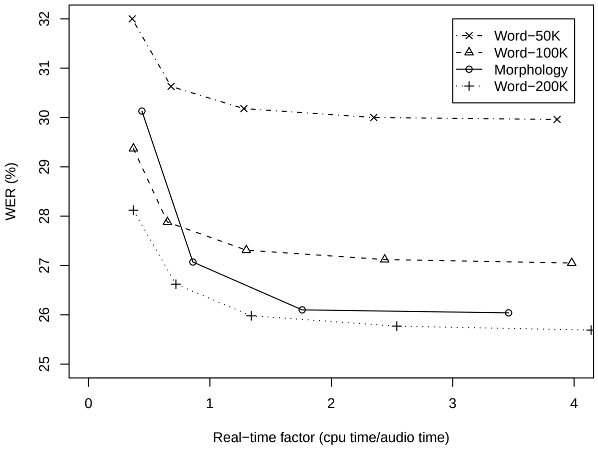

Fig. 3. Word error rate versus real-time factor obtained by changing the pruning beam width

Note that the word-based models output word sequences such as *"merhaba saat on üç haberleri ajanstan alıyorsunuz"*, while the morphology-based model outputs lexical morpheme sequences such as *"merhaba[Noun] saat[Noun] on[Adj] üç[Adj] haber[Noun] +lAr[A3pl] +SH[P3sg] ajans[Noun] +DAn[Abl] al[Verb] +Hyor[Prog1] +sHnHz[A2pl]"*. Therefore, we use the morphological parser as a word generator to convert the recognition output to words.

Figure 3 shows the word error rate versus run-time factor for the baseline word models and the morphology-integrated model. The morphology-integrated model outperforms the 50K and 100K word models while the 200K word model is slightly better.

V. DISCUSSION

The experimental results prove that integrating the morphology into the search network is very effective in reducing OOV rates with relatively fewer number of lexical units and thus improving the recognition accuracy. While the 200K vocabulary word model performs slightly better than the morphology model, the computational and memory requirements for building and optimizing this model is greatly higher than constructing the morphology-integrated model. Even if the morphology-integrated model can not be fully optimized, the local determinization seems to work well in practice.

The major advantage of the morphology model is that it can recognize infinite number of words, while the static vocabulary word models are limited to words in the vocabulary and the training corpus. It also outputs only grammatical words in contrast to sub-word based models.

The morphology-integrated model can be improved by using better language models over lexical morphemes. Since this model outputs the morphological parse of the words annotated with morphological features, this morphological information can be exploited in a feature-based language model. For instance, the morpheme lattices output from the decoder can be rescored with a Maximum Entropy feature based language model or a discriminative language model as in [16].

ACKNOWLEDGMENT

This work was supported by the Boğaziçi University Research Fund under the grant numbers 06A102 and 08M103, the Scientific and Technological Research Council of Turkey (TÜBİTAK) under the grant number 107E261, the Turkish State Planning Organization (DPT) under the TAM Project, number 2007K120610 and TÜBİTAK BİDEB 2211.

REFERENCES

[1] T. Hirsimäki, M. Creutz, V. Siivola, M. Kurimo, S. Virpioja, and J. Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to finnish," *Computer Speech & Language*, vol. 20, no. 4, pp. 515–541, 2006.

[2] E. Arısoy, D. Can, S. Parlak, H. Sak, and M. Saraçlar, "Turkish broadcast news transcription and retrieval," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 5, pp. 874–883, 2009.

[3] H. Erdogan, O. Buyuk, and K. Oflazer, "Incorporating language constraints in sub-word based speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2005.

[4] E. Arısoy and M. Saraçlar, "Lattice extension and vocabulary adaptation for turkish lvcsr," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 1, pp. 163–173, 2009.

[5] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech & Language*, vol. 16, no. 1, pp. 69–88, 2002.

[6] H. Sak, T. Güngör, and M. Saraçlar, "Turkish language resources: Morphological parser, morphological disambiguator and web corpus," in *GoTAL 2008*, ser. LNCS, vol. 5221. Springer, 2008, pp. 417–427.

[7] K. Koskenniemi, "A general computational model for word-form recognition and production," in *ACL*, 1984, pp. 178–181.

[8] K. Oflazer, "Two-level description of Turkish morphology," *Literary and Linguistic Computing*, vol. 9, no. 2, pp. 137–148, 1994.

[9] L. Karttunen and K. R. Beesley, "Two-level rule compiler," Xerox Palo Alto Research Center, Palo Alto, CA, Tech. Rep., 1992.

[10] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *CIAA 2007*, ser. LNCS, vol. 4783. Springer, 2007, pp. 11–23, http://www.openfst.org.

[11] K. Oflazer and S. Inkelas, "The architecture and the implementation of a finite state pronunciation lexicon for Turkish," *Computer Speech and Language*, vol. 20, no. 1, pp. 80–106, 2006.

[12] M. Mohri and M. Riley, "Integrated context-dependent networks in very large vocabulary speech recognition," in *Eurospeech*, 1999, pp. 811–814.

[13] C. Allauzen, M. Mohri, M. Riley, and B. Roark, "A generalized construction of integrated speech recognition transducers," in *ICASSP*, 2004.

[14] C. Allauzen, M. Mohri, and B. Roark, "The design principles and algorithms of a weighted grammar library," *International Journal of Foundations of Computer Science*, vol. 16, no. 3, pp. 403–421.

[15] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proceedings of ICSLP*, vol. 2, 2002, pp. 901–904.

[16] B. Roark, M. Saraçlar, and M. Collins, "Discriminative n-gram language modeling," *Computer Speech and Language*, vol. 21, no. 2, pp. 373–392, April 2007.