MORPHOLOGY-BASED AND SUB-WORD LANGUAGE MODELING FOR TURKISH SPEECH RECOGNITION

Haşim Sak [1], *Murat Saraçlar* [2], *Tunga Güngör* [1]

[1] Computer Engineering, Boğaziçi University, Bebek, İstanbul, Turkey
[2] Electrical & Electronics Engineering, Boğaziçi University, Bebek, İstanbul, Turkey
{hasim.sak, murat.saraclar, gungort}@boun.edu.tr

ABSTRACT

We explore morphology-based and sub-word language modeling approaches proposed for morphologically rich languages, and evaluate and contrast them for Turkish broadcast news transcription task. In addition, as a morphology-based model, we improve our previously proposed *morphology-integrated* model for automatic speech recognition. This model is built by composing the finite-state transducer of the morphological parser with a language model over lexical morphemes. This approach provides a morphology-integrated search network with an unlimited vocabulary, generating only valid word forms while reducing the out-of-vocabulary rate and hence improving the word error rate. We also analyze the effect of morphotactics and morphological disambiguation on the speech recognition accuracy for the morphology-integrated model. The improved morphology-integrated model performs better than statistically derived sub-word models with added benefit of generating morphosyntactic and semantic features.

Index Terms— Morphology-based, sub-word, morphology-integrated, language modeling, automatic speech recognition.

1. INTRODUCTION

Morphologically rich languages such as Turkish, Finnish, and Arabic present some challenges for language modeling. Such languages have relatively high number of out-of-vocabulary (OOV) words due to rapid vocabulary growth, leading to higher word error rates (WERs) in automatic speech recognition (ASR). Moreover, having a large number of words causes data sparseness and leads to non-robust parameter estimates for n-gram language models. Turkish, being an agglutinative language with a highly productive inflectional and derivational morphology is especially prone to these problems.

A commonly used approach to reduce OOV rate and alleviate data sparsity is using sub-word units for language modeling [1], [2], [3]. Sub-word units can be grammatical units such as morphemes, or some grouping of them such as stem and ending (grouping of suffixes). They can also be obtained by splitting words into morpheme like units using unsupervised statistical methods. Sub-word language models alleviate the OOV problem, however the speech decoder can generate ungrammatical sub-word sequences and post-processing of sub-word lattices can increase the recognition accuracy [3], [4].

Factored Language Models (FLMs) with a generalized backoff mechanism have been proposed as a morphology-based language modeling technique to improve the robustness of probability estimates for rarely observed word n-grams [5]. While FLMs are not used to deal with OOV problem, they have been shown to be effective in reducing language model perplexity thanks to better smoothing.

We proposed a novel approach for building a morphology-integrated model for ASR in morphologically rich languages [6]. We built this model by composing the finite-state transducer for morphological parser with a language model over lexical morphemes. This model has the advantage of the dynamic vocabulary in contrast to word models and it only generates valid word forms in contrast to sub-word models.

In this paper, we compare and contrast several morphology-based and sub-word language modeling approaches for Turkish broadcast news transcription task. The results are given for both first-pass recognition and lattice rescoring. Moreover, we improve the performance of the morphology-integrated model by automatically expanding the root lexicon of the parser. We also analyze the effect of morphological disambiguation and the morphotactics (morpho-syntax) on the morphology-integrated model.

2. LANGUAGE RESOURCES

The text corpora that we used for estimating the parameters of statistical language models are composed of 182.3 million-words BOUN NewsCor corpus collected from news portals in Turkish [7] and 1.3 million-words text corpus (BN Corpus) obtained from the transcriptions of the Turkish Broadcast News speech database [2].

For morphological analyses, we use our finite-state transducer implementation of the morphological parser, which is obtained as the composition of the morphophonemics transducer encoding the phonological rules such as for vowel harmony phenomena, and the morphotactics transducer encoding the morphosyntax of the language. The lexicon of the parser contains 55,278 root words. 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 1 shows the part of the lexical transducer of this modified morphological parser for the word *haberleri*. The morphological features are identical to the ones in [8]. As can be seen from this figure, there is ambiguity in morphological parsing. We used the averaged perceptron-based morphological disambiguator [7] to resolve this ambiguity. The disambiguation system achieves about 97.05% disambiguation accuracy on the test set.
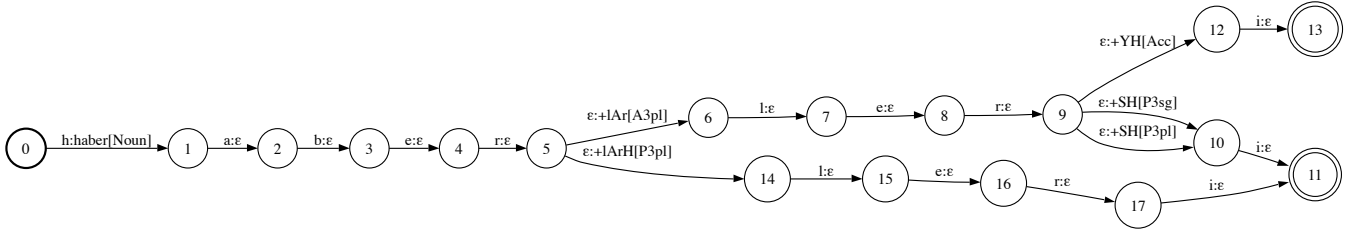
Fig. 1. Part of the lexical transducer of the parser representing the ambiguous analyses for the word *haberleri*

3. LANGUAGE MODELS

In the following sections, we describe the word, sub-word and morphology-based language models. The corresponding statistical and grammatical splitting approaches are shown for an example sentence in Figure 2.

3.1. Words

The conventional approach for language modeling is estimating a statistical *n*-gram language model over a fixed vocabulary of words. The problem for agglutinative languages is having relatively higher number of word types, resulting in higher OOV rates in a test set, hence increasing WER of the ASR system. The high number of language modeling units also causes data sparsity problem resulting in non-robust parameter estimates. Increasing the vocabulary size reduces the OOV rate, however it also increases the computational requirements of the system and requires more data for robust parameter estimation. In this regard, Turkish is especially challenging since it has a very productive morphology with theoretically infinite number of words. As a baseline word language model, we built 200K vocabulary *3*-gram language model. Our previous study showed that higher vocabulary sizes than 200K and higher *n*-gram orders did not improve the accuracy significantly [2]. The OOV rate for 200K word vocabulary is about 2% on the test set.

3.2. Statistically Derived Sub-words (Morphs)

For unlimited vocabulary speech recognition, splitting words into morphs (morpheme-like sub-words) using an unsupervised algorithm based on Minimum Description Length principle has been very effective by alleviating OOV problem and reducing language model perplexity [1]. The baseline algorithm introduced in [9] is used to segment word types in the text corpus. We used the best performing segmentations of the study in [2]. This segmentation results in 76K morphs types where the non-initial morphs are marked with "+" symbol to facilitate the restoration of word boundaries in speech recognition output. The OOV rate for the statistical morphs model is 0% on the test set, since the letters are also included in the morphs lexicon. For ASR experiments, the optimal *n*-gram order was chosen as 4. Statistical morphs have the advantage that no linguistic knowledge is required about the language. On the other hand, since morphs do not generally correspond to grammatical morphemes, we can not easily employ linguistic information in later stages of processing such as rescoring sub-word lattices. Moreover, 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].

3.3. Lexical Morphemes

We previously proposed a more grammatical approach for incorporating the morphological information into the language model [6]. In this approach, the finite-state transducer for the morphological parser replaces the static lexicon as a dynamic computational lexicon. Also an *n*-gram language model over lexical morphemes, which are the output units of the computational lexicon, is estimated over a morphologically parsed and disambiguated text corpus. The composition of this lexicon with the language model gives us a morphology-integrated search network for ASR.

The OOV rate for the morphological parser is about 3% on the test set. We expanded the root lexicon of the morphological parser to reduce the OOV rate to better compete with the statistical morphs. The most frequent 20K unparsed words in the training corpus, which are mostly common misspellings and proper nouns, are added to the root lexicon. The most frequent unparsed 30K words with an apostrophe, which are mostly proper nouns and some misspellings of proper nouns, are also added to the lexicon. The OOV rate of the parser is effectively reduced to 0.68% on the test set. For ASR experiments, the optimal *n*-gram order was chosen as 4.

We expect the morphological parser to be effective in constraining the language model over lexical morphemes by allowing only valid morpheme sequences when composed with the language model thanks to the morpho-syntax of the parser. To test this in ASR experiments we modified the morphotactics component of the morphological parser and allowed any morpheme sequences. We also experimented with the effect of the morphological disambiguation on ASR performance. To experiment with this, we did not disambiguate the text corpora with the perceptron-based morphological disambiguator. Rather, we selected the morphological parse with the least number of morphemes for each word.

3.4. Factored Language Models

As another morphology-based language modeling technique, *Factored Language Models (FLMs)* have been shown to reduce language model perplexity and lead to WER reductions in speech recognition systems [5]. FLMs decompose words into a set of features (or factors) and estimate a language model over these factors, smoothed with *generalized parallel backoff* mechanism which improves the robustness of probability estimates for rarely observed *n*-grams.

FLMs can use factors representing morphological, syntactic, or semantic word information. For Turkish, we used the lexical word (*W*), the stem of the word (*S*), the lexical ending (*E*), the last morpheme (*M*) and the part of speech tag (*T*) of the word. For instance, the morphological analysis *haber[Noun]+lAr[A3pl]+SH[P3sg]* of the word *haberleri* is factored as follows:

W-haber[Noun]+lAr[A3pl]+SH[P3sg]:*S*-haber[Noun]:

gloss: hello you are getting the news from the agency
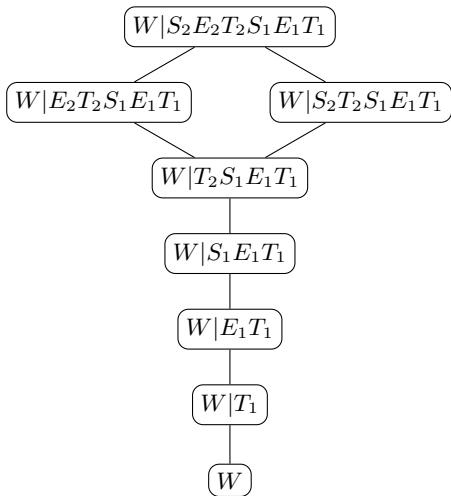
words: merhaba haberleri ajanstan alıyorsunuz

statistical morphs: merhaba haber +ler +i ajans +tan al +ıyor +sun +uz

lexical morphemes: merhaba[Noun] haber[Noun] +lAr[A3pl] +SH[P3sg] ajans[Noun] +DAn[Abl] al[Verb] +Hyor[Prog1] +sHnHz[A2pl]

Fig. 2. Statistical and grammatical word splitting approaches.

E-+lAr[A3pl]+SH[P3sg]:*M*-+SH[P3sg]:*T*-[Noun]

Although we experimented with the genetic algorithms program for automatically learning FLM backoff graph structures [5], we could not find a better backoff graph structure than our hand-selected one. The best performing hand-selected backoff graph is shown in Figure 3. The graph is linear except at the first backoff step. The last morpheme factor *M* was not used for this backoff graph. The first backoff step combines two backoff path estimates which drop the oldest stem and ending. Then, we proceed by dropping the furthest factor at each step.

$$W|S_2 E_2 T_2 S_1 E_1 T_1$$

$$W|E_2 T_2 S_1 E_1 T_1 \qquad W|S_2 T_2 S_1 E_1 T_1$$

$$W|T_2 S_1 E_1 T_1$$

$$W|S_1 E_1 T_1$$

$$W|E_1 T_1$$

$$W|T_1$$

$$W$$

Fig. 3. The best performing hand-selected back-off graph for factored language models.

The factored NewsCor corpus contains about 2M lexical words and the OOV rate is 0.84% on the test set. To limit the computational requirements, we constrained the vocabulary size to 285K giving an OOV rate of about 2% on the test set, which can be achieved with the 200K vocabulary word model. The perplexity values for standard word LM, the lexical word LM (*W* factor) and the FLM of Figure 3 are given in Table 1 for the small 1.3M words BN Corpus and the large 200M words NewsCor Corpus.

Table 1. Word LM and FLM perplexities of BN Corpus/NewsCor Corpus on the test set

Model	BN OOV (%)	BN Perpl.	NewsCor OOV (%)	NewsCor Perpl.
Word	4.4	644	2	523
Lexical Word	5	668	2	579
FLM	5	608	2	556

4. ASR EXPERIMENTS

We evaluated the performance of morphology-based and sub-word language models on a Turkish broadcast news transcription task.

The acoustic model uses hidden Markov models (HMMs) trained on 188 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 and 23,038 words). We used the geometric duration modeling in the decoder.

The weighted finite-state transducers (WFSTs) are used for representing all knowledge sources in the ASR system [10]. 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. However the morphology-integrated model can not be fully optimized, since the finite-state transducer of the morphological parser is cycle-ambiguous and can not be determinized. Still, we can apply the local determinization algorithm for locally optimizing the search network using the *grm-localdeterminize* utility from *AT&T Grammar Library* [1]. The other search networks are constructed by composing and optimizing lexicon and grammar composition $L \circ G$ with the *dmake* utility from *AT&T DCD Library* [2]. The context-dependency and the hidden Markov models are composed on-the-fly in the decoder. All language models were estimated by linearly interpolating two language models trained over the BOUN NewsCor corpus and the BN Corpus to reduce the effect of out-of-domain data using the SRILM toolkit [3].

Figure 4 shows the word error rate versus run-time factor for the proposed morphology-integrated model *MP*, 200K vocabulary word model *Word-200K*, morphology-integrated model with reduced OOV *MP-OOV*, statistical morphs model *Morphs*, MP model with no disambiguation *MP-OOV-NoDisamb*, and MP model without proper morphotactics *MP-OOV-NoMT*. The difference between statistical morphs model and the improved morphology-integrated models are not statistically significant.

Since the language models are pruned for computational reasons when building the optimized search networks of first-pass recognition, we rescore the output lattices with unpruned language models as a second-pass. The lattice rescoring experiments were carried out only for the lattices produced with the prune beam width 12. The WERs for the first pass and after rescoring are given in Table 2. We also experimented with *n*-best list rescoring of the rescored lattice output from the morphology-integrated model with the FLM described in section 3.4. However, we could not achieve any improvements possibly due to higher OOV rates and computational limitations of the FLM model.
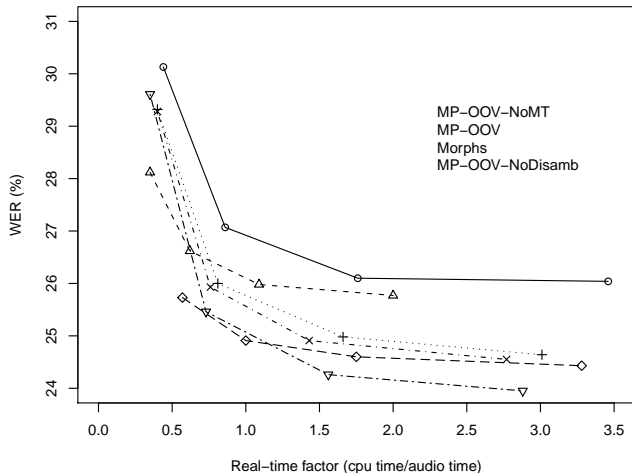
Fig. 4. Word error rate versus real-time factor obtained by changing the pruning beam width from 9 to 12.

Table 2. Rescoring Results

Model	WER (%)	Rescore WER (%)
Word	25.77	24.55
MP	24.55	22.84
Morphs	24.43	22.66
MP No Disamb	23.95	22.50

5. DISCUSSION

As seen in Figure 4, the improvement in OOV rate of the morphological parser translates to WER reductions bringing the accuracy of the morphology-integrated model very close to the statistical morphs model. Since we can only locally determinize the morphology-integrated models, they significantly perform worse for small real-time factors. The morphotactics also has very negligible effect in constraining the language model and improving the language model performance. As an interesting observation, choosing the morphological parse with the least number of morphemes is more effective than morphological disambiguation of the training corpus. This can be explained by the fact that ambiguous parses increase the data sparsity, therefore it is better to choose the same morphological parse of a word for every context. The lattice rescoring experiments show that disambiguation effect is probably only significant for pruned language models.

We have observed that FLMs are more effective in reducing perplexity of language models when the training data is limited. When the training data size increases, we have more robust parameter estimates, therefore limiting the advantage of the FLMs. FLMs are also computationally more expensive as the data size increases. Another disadvantage over standard n-grams is that they can not be easily used in the first-pass of the conventional speech recognizers, since they can not be efficiently and compactly represented as finite-state models. While they can be used to estimate word n-gram probabilities, they require explicit representation of all n-grams in finite-state model representation, limiting this approach only to moderate vocabulary sizes. Since agglutinative languages have a large number of word types, and FLMs do not aim to reduce OOV rate, FLMs are more useful when the training data is limited.

Morphology-integrated models also have the advantage that lexical morpheme lattices carry explicit linguistic information. Therefore, this approach can present useful morpho-syntactic and morpho-semantic features for rescoring with Maximum Entropy models [11] and discriminative language modeling [12], [2].

6. REFERENCES

[1] Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen, "Unlimited vocabulary speech recognition with morph language models applied to Finnish," *Computer Speech & Language*, vol. 20, no. 4, pp. 515–541, 2006.

[2] Ebru Arısoy, Doğan Can, Sıddıka Parlak, Haşim Sak, and Murat Saraçlar, "Turkish broadcast news transcription and retrieval," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 5, pp. 874–883, 2009.

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

[4] Ebru Arısoy and Murat 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] Katrin Kirchhoff, Dimitra Vergyri, Jeff Bilmes, Kevin Duh, and Andreas Stolcke, "Morphology-based language modeling for conversational Arabic speech recognition," *Computer Speech & Language*, vol. 20, no. 4, pp. 589–608, 2006.

[6] Haşim Sak, Murat Saraçlar, and Tunga Güngör, "Integrating morphology into automatic speech recognition," in *ASRU*, 2009, Accepted.

[7] Haşim Sak, Tunga Güngör, and Murat Saraçlar, "Turkish language resources: Morphological parser, morphological disambiguator and web corpus," in *GoTAL 2008*. 2008, vol. 5221 of *LNCS*, pp. 417–427, Springer.

[8] Kemal Oflazer and Sharon 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.

[9] Mathias Creutz and Krista Lagus, "Unsupervised discovery of morphemes," in *Proceedings of the Workshop on Morphological and Phonological Learning of ACL*, 2002, pp. 21–30.

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

[11] Ruhi Sarikaya, Mohamed Afify, Yonggang Deng, Hakan Erdogan, and Yuqing Gao, "Joint morphological-lexical language modeling for processing morphologically rich languages with application to dialectal Arabic," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 7, pp. 1330–1339, Sept. 2008.

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