

A COMPARISON OF FOUR LANGUAGE MODELS FOR LARGE VOCABULARY TURKISH SPEECH RECOGNITION

Helin Dutagaci, Levent M. Arslan

Department of Electrical and Electronics Engineering
Bogazici University
[dutagach, arslanle]@boun.edu.tr

ABSTRACT

This paper gives a comparison of three language models proposed as alternatives to word-based language model for large vocabulary speech recognition of Turkish. Turkish is an agglutinative language and has morphological productivity. This results in a huge vocabulary size and a large number of out of vocabulary words for unseen test data. The solution is to parse the words, in order to get smaller base units, which are capable of covering the language with relatively small vocabulary size. Three different ways of decomposing words into base units are described: Morphem-based model, stem-ending-based model and syllable-based model. These models are compared with respect to vocabulary size, coverage, number of out of vocabulary words, perplexity and sensitivity to context. For all three models, a significant improvement for those measures are observed compared to the word-based language model.

1. INTRODUCTION

Large vocabulary speech recognition systems require a vocabulary that covers an acceptable portion of the language. This requirement becomes challenging for Turkish if words are used as base units, since Turkish is an agglutinative language. The morphological structure of Turkish makes it possible to derive thousands of different words from a single stem.

For other languages that are also morphologically productive or compounding such as German, Portuguese, Swedish, models based on morphological decomposition or compound-splitting were proposed [1,2,3]. For Turkish, there are current research efforts on computational linguistics and natural language processing but very little work is done for language modeling with the purpose of speech recognition [4,5].

In this paper, as alternatives to word-based model, three language models which are named according to their base units are described: Morphem-based model, stem-ending based model and syllable-based model. The training and test text corpora were each parsed according to these models and tests were performed in order to measure the appropriateness of each model to large vocabulary speech recognition task.

In section 2, the morphem-based, stem-ending-based and the syllable-based models are defined. Section 3 gives the properties of the training and test database. In section 4, the statistics and test results are given for each language model,

and are discussed. In section 5, conclusion and future work are given.

2. TURKISH MORPOLOGY AND THE LANGUAGE MODELS

Turkish words gain different meanings and functionalities as they are concatenated with suffixes one after another. Kemal Oflazer gave a two-level morphological description for Turkish [6]. He explained the morphotactics of Turkish as finite-state. There are a number of phonetic rules (vowel harmony, deletion, insertion, etc.) that construct the surface realizations of morphem additions.

A morphological parser was written in Prolog at Bogazici University Computer Science Department. It works as a finite-state transducer, using the finite-state structure and the phonetic rules described by Oflazer. The parser utilizes a dictionary of 29541 stems and 111 suffixes (corresponding to 479 different surface realizations).

As examples to the morphological productivity of Turkish, following words, with common Turkish stems, are given. The words *gel* and *güzel* are stems. The “-”s are used to indicate the morphem boundaries.

<i>gel</i>	<i>(come)</i>
<i>gel-iyor</i>	<i>(he/she is coming)</i>
<i>gel-iyor-du-m</i>	<i>(I was coming)</i>
<i>güzel</i>	<i>(beautiful)</i>
<i>güzel-leş</i>	<i>(become beautiful)</i>
<i>güzel-leş-me-z-se-k</i>	<i>(if we do not become beautiful)</i>

Morphem-based model considers morphems (stems and suffixes) as units. For the examples above units *gel*, *iyor*, *du*, *m*, *güzel*, *leş*, *me*, *z*, *se* and *k* are considered as separate units. The advantage of the morphem-based model is its ability to cover new words constructed by derivations and inflections. The disadvantage is that the base-units are short, reducing acoustic differentiability of base-units.

To overcome the disadvantage of short base units stem-ending-based model can be utilized. Stem-ending model is also a morphology-based model. A word is considered to have two parts; the stem and all suffixes added to the stem as a whole (which is named as “ending”). Therefore, our units for the above words are *gel*, *iyor*, *iyordum*, *güzel*, *leş* and *leşmezsek*. The drawback with the stem-ending-based model is the vast amount of possible endings in Turkish. Number of

suffixes added to a stem can be up to 10 or more; and number of suffix combinations is large.

Syllable-based model is the simplest one; since it does not require morphological knowledge. A syllable in Turkish cannot be of more than 4 phonemes. Each syllable contains exactly one vowel, which makes it suitable as acoustic base unit.

3. DATABASE

A training text corpus consisting of 1,1 million words was constructed using online books. Most of these books were from literature and social sciences. We used a set of test texts from different domains, some of which were not represented in the training corpus. Training and test texts were morphologically decomposed to form morphem-tokenized and stem-ending-tokenized versions.

In morphem-tokenized texts, the lexical descriptions of the suffixes were used instead of their surface realizations. Since surface realizations of suffixations in Turkish are deterministic, they will have no contribution to the statistics of the language model.

	word-based	morphem-based	stem-ending-based	syllable-based
# tokens	1,116,451	2,217,534	1,785,934	3,027,791
tokens per word	1	2.0	1.6	2.7

Table 1: Number of tokens in training data

For the stem-ending-tokenized texts it was more appropriate to use surface realizations. Undecomposed words (words that were not recognized by the parser) were left in their place as words with single morphem. Of the 1,116,451 words sent to the parser, 1,048,074 (94%) were decomposed and 6% remained undecomposed. Table 1 shows the number of tokens and the average number of tokens per word for all four models.

	# words	percentage of undecomposed	representation in training
narrations	50,088	2.3	yes
TV news	91,451	9.3	no
constitution	18,210	1.6	no
fiction 1 (translation)	64,331	5.5	yes
fiction 2 (translation)	22,904	5.9	yes
native fiction	20,949	2.4	yes
psychiatry	26,289	11.8	no
total-test	294,222	6.3	-

Table 2: Properties of test data

The average number of morphemes in a word is estimated as 2. Of the decomposed words in the training data 36.1% are in the stem form and 99.4% of the words have less than 6 morphemes. In the training data, the maximum number of morphemes in a word was 10. When we also count the undecomposed words, 40% of the words have no ending.

The maximum number of syllables in a word is 13 for the training data. Words with more than 6 syllables are not very typical. When we compare the histograms of morphem and syllable numbers of words we see that they are similar except for the first two counts. This is because a typical stem in Turkish consists of two syllables.

There are 7 test texts from different domains. Table 2 shows their properties; number of words, percentage of morphologically undecomposed words, whether or not the domain is represented in the training set.

Percentages of morphologically undecomposed words for the TV news and psychiatry are high since TV news contains lots of proper nouns and in psychiatric articles terms not present in the parser’s dictionary are frequently used.

4. MEASUREMENTS ON THE MODELS

The “CMU-Cambridge Toolkit” [7] is used to estimate the number of OOV words and perplexity values. Size of the vocabularies used for those estimations, number of distinct tokens in each training text tokenized according to the corresponding language model is shown in Table 3. For word-based model the vocabulary is limited to the most frequent 60K words because of computational limitations.

	word-based	morphem-based	stem-ending-based	syllable-based
# tokens	1,116,451	2,217,534	1,785,934	3,027,791
# distinct tokens	144,986	34,830	51,634	5,243
vocabulary size	60,000	34,830	51,634	5,243

Table 3: Number of distinct tokens in training data

The training corpus contains 144,986 distinct tokens. With a 60K vocabulary, which has already large size, the training data cannot be covered completely. Without processing the test data, it is obvious that the word-based model promises high number of out of vocabulary words. Also, this fact leads to very coarse bigram and trigram probability estimations; since a training data with about 1M words is far from sufficient to correctly estimate the statistics of 144,986 tokens or 60K words of the vocabulary we selected.

The number of distinct endings occurred in the stem-ending-tokenized training text is 17,368. This number demonstrates the morphological productivity of Turkish.

Figure 1 and Figure 2 shows the coverage of training and test data with respect to vocabulary size. From the graphs we can conclude that the syllable-based and morphem-based models catch high levels of coverage of both training and test

data with relatively small vocabularies. The morphem-based model covers more data than all other models with vocabulary sizes less than 1K. Actually the most frequent 1000 tokens of Turkish correspond to suffixes and most common stems. On the other hand the word-based model does not show any sign of saturation.

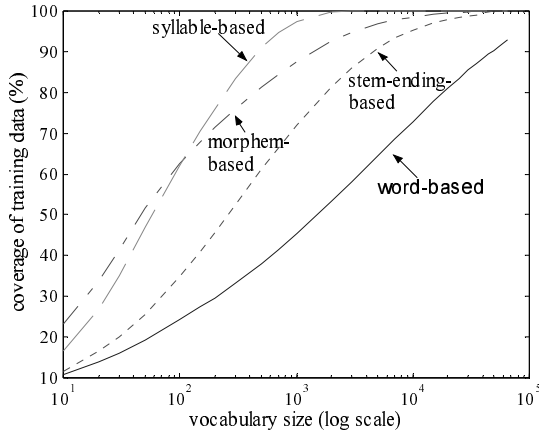


Figure 1: Coverage of training data

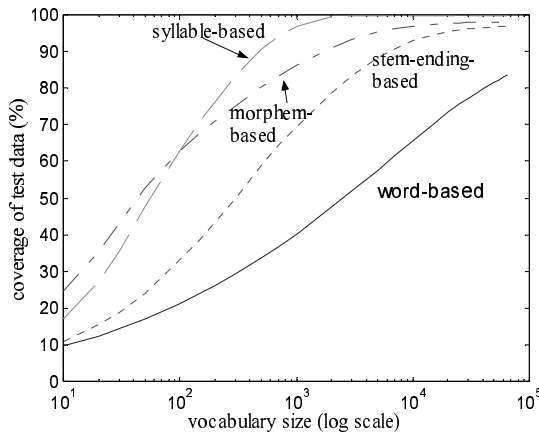


Figure 2: Coverage of test data

Table 4 shows percentage of out of vocabulary words for the training text and test texts, with the vocabulary sizes (indicated in parenthesis) used for each language model. Out of the three domains (TV news, constitution and psychiatry) that were not represented in the training data, TV news and psychiatry texts have high out of vocabulary percentage.

The constitution of Turkish Republic, is a short text, it contains fewer proper nouns and domain specific terms. For the TV news and the psychiatry texts, one reason for the high OOV percentage is the unseen proper nouns and terminology; but that fact does not explain the 28%, since one word in four can not be of domain terminology. That is related to expressions specific to that domain, which are morphologically produced from ordinary stems. The morphem-based model resolved those expressions and

resulted in 5.36% OOV words, which are assumed to be domain specific terms.

There is a dramatic reduction in the percentages of OOV words for the other three models compared to the word-based model. Although the OOV percentages for stem-ending-based are highest among those three, it yields a significant improvement over the word-based model. However, the stem-ending-based model has to utilize larger vocabularies to achieve an improvement comparable to morphem-based and stem-ending-based models.

Another drawback with stem-ending-based model is, although it contained 17K endings in its vocabulary, it does not cover all the possible endings in an unseen Turkish text. In this sense, the morphem-based model seems more flexible, since it can catch any morphological expression possible in Turkish as far as the stem exists in the vocabulary.

	word-based (60K)	morphem-based (34,830)	stem-ending-based (51,634)	syllable-based (5,243)
training	7.92	0	0	0
narrations	15.37	0.98	1.76	0.02
TV news	18.94	2.68	4.01	0.24
constitution	12.24	0.54	1.18	0.01
fiction 1 (translation)	14.50	1.59	2.54	0.04
fiction 2 (translation)	14.21	1.97	2.94	0.40
native fiction	14.82	1.04	1.53	0.02
psychiatry	27.87	5.36	7.95	1.24

Table 4: OOV percentages of training and test texts

Comparing Table 2 and Table 4, we see that the OOV percentages are lower than the percentages of morphologically undecomposed words. The dictionary of the parser is fixed, and it contained only stems and suffixes. The vocabulary obtained from the morphem-tokenized training data contains 5000 words more than the parser's dictionary and it is data-driven. It contains frequent proper nouns, frequent terms, words whose stems are not present in the parser's dictionary.

The syllable-based model resulted in fewer OOV words with a small vocabulary, since it utilizes the shorter elements of the language as base units. This model has very little sensitivity to domain, since its base-units do not have meanings. The morphem-based model also has small units (i.e., the suffixes of Turkish) in its vocabulary; but the majority of its units are stems. That is an important advantage of morphem-based model over syllable-based model, since recognizing the stem correctly has crucial importance. Errors in the suffixation can be corrected by more sophisticated algorithms considering context in the sentence level but when

the stem is not correctly recognized it is more difficult to recover it.

Table 5 shows the perplexity values of trigram models with respect to training and test data. We observe very high perplexity for word-based model. One reason is the large vocabulary size; second and not less significant reason is the free word order in Turkish. This fact is demonstrated with about 35% bigram hits and 11% trigram hits of the test data.

The syntax of Turkish is much more flexible than that of English.

Perplexity values with syllable-based model are the lowest. There are three reasons for this. First, the vocabulary is small. Second, the probability of a particular sequence of two or three syllables (since a word contains them in the same order!) is higher than the probability of a bigram or trigram of words. Third reason is the strong vowel harmony in Turkish words.

The morphem-based model yields low perplexity values compared to word-based and stem-ending-based models. It should be noted that it is less sensitive to context. For the three test texts that were not represented in the training data, the perplexities were much higher than others for word-based and stem-ending-based models. For morphem-based model, the differences were not as high.

	word-based (60K)	morphem-based (34,830)	stem-ending-based (51,634)	syllable-based (5,243)
training	41	26	27	17
narrations	2313	108	354	47
TV news	6484	113	520	42
constitution	4861	93	395	30
fiction 1 (translation)	3306	87	314	37
fiction 2 (translation)	2939	90	295	43
native fiction	5438	140	488	54
psychiatry	5300	100	525	61

Table 5: Perplexities of trigram models

5. CONCLUSION AND FUTURE WORK

When we compare the four models we can conclude at the following: Using small base units results in a disadvantage. The function of a language model, with its n-gram probabilities, is to search for meaningful units that best fit the given context. The syllable and morphem-based models deal with parts of the words, and take them out of the context they were used. The stem-ending-model has the advantage of representing stems in their context with a trigram model.

On the other hand, morphems in Turkish corresponds to functional words (which dominate on bigrams and trigrams) of English. The functions handled by syntax in English are handled by morphology in Turkish. When we look at the

most common words in English, we see that they correspond to suffixes of Turkish. Consider the following Turkish word:

uygar-laş-tır-ama-dık-lar-ımız

(those we could not make become civilized)

This is the reason why perplexities with the morphem-based model are closest to that of English.

The morphem-based and stem-ending-based models have the advantage of using stems as the base units. The base units of the syllable-based model do not have meanings. Another important issue is the ability of morphem-based and stem-ending-based models to detect the word boundaries, which becomes a challenging task with the syllable-based model.

When complexity is not an issue, morphem-based and stem-ending-based models are more suitable for speech recognition applications rather than the syllabus-based model.

We are now constructing the networks for recognition experiments. Morphem-based model demands a complex, finite-state network, with its surface realizations as states. Networks for stem-ending-based and syllable-based models will be relatively simple. Word error recognition rates will measure the appropriateness of each base unit selection. We also plan to increase size of the corpus up to 10 million words to derive more accurate statistics of the language. But we don't expect the properties of the four models will change too much relative to each other.

6. REFERENCES

- [1] Geutner, P., "Using Morphology Towards Better Large-Vocabulary Speech Recognition Systems", *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-1995*, vol.1, pp. 445-448, 1995.
- [2] Martins, C., Neto, J.P. and Almeida, L., "Using Partial Morphological Analysis in Language Modeling Estimation for Large-Vocabulary Portuguese Speech Recognition", *Proceedings EUROSPEECH 99*, Budapest, Hungary, 1999.
- [3] Carter, D. and al., "Handling Compound Nouns in a Swedish Speech-Understanding System", in *Proceedings of ICSLP 96*, Philadelphia, USA, 1996.
- [4] Çarkı, K., Geutner, P. and Schultz, T., "Turkish LVCSR: Towards Better Speech Recognition for Agglutinative Languages", *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-2000*, Istanbul, Turkey, June 2000.
- [5] Mengüşoğlu, E. and Deroo, O., "Turkish LVCSR: Database preparation and Language Modeling for an Agglutinative Language", *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-2001 Student Forum*, Salt Lake City, May 2001.
- [6] Oflazer, K., "Two-level Description of Turkish Morphology", *Proceedings of the 6th Conference of the European Chapter of the ACL*, April 1993.
- [7] Clarkson, P.R. and Rosenfeld, R., "Statistical Language Modeling Using the CMU-Cambridge Toolkit", *Proceedings of EUROSPEECH 97*, Rhodes, Greece, 1997.