Building and Incorporating Language Models for Persian Continuous Speech Recognition Systems

M. Bahrani, H. Sameti, N. Hafezi and H. Movasagh
Speech Processing Lab, Computer Engineering Department,
Sharif University of Technology, Tehran, Iran
bahrani@ce.sharif.edu, sameti@sharif.edu, hafezi@ce.sharif.edu, hmovasagh@yahoo.com

Abstract

In this paper building statistical language models for Persian language using a corpus and incorporating them in Persian continuous speech recognition (CSR) system are described. We used Persian Text Corpus for building the language models. First we preprocessed the texts of corpus by correcting the different orthography of words. Also, the number of POS tags was decreased by clustering POS tags manually. Then we extracted word based monogram and POS-based bigram and trigram language models from the corpus. We also present the procedure of incorporating language models in a Persian CSR system. By using the language models 27.4% reduction in word error rate was achieved in the best case.

Keywords: statistical language models, Continuous Speech Recognition, Persian Text Corpus.

1. Introduction

For recognizing continuous speech, the acoustic signal is too weak to narrow down the number of word candidates. Hence, speech recognizers employ a language model that prunes out acoustic alternatives by taking into account the previous words that were recognized. N-gram language models are frequently used by the large vocabulary speech recognition systems to constrain and guide the search (X. Huang et al., 1993; S. J. Young et al., 1995).

The speech recognition problem is viewed as finding the most likely word sequence given the acoustic signal (L. Rabiner & B.H. Juang, 1993):

$$\begin{aligned}\hat{W} &= \arg\max_{W} P(W \mid A) \\ &= \arg\max_{W} P(A \mid W).P(W)\end{aligned} \qquad (1)$$

The last line involves two probabilities that need to be estimated, the first due to the acoustic model $P(A \mid W)$ and the second due to the language model $P(W)$. In this paper, $P(W)$ is estimated with monogram, bigram and trigram language models.

To build statistical language models, a large amount of training data is required. Unfortunately, collecting data and providing a suitable text corpus for Persian language is in its primary stages, hence a few studies have been done for incorporating language models in Persian speech recognition systems.

In this paper we have used "Persian Text Corpus", the only text corpus in Persian, to build statistical language models, and then we have incorporated those models in a Persian continuous speech recognition system.

In section 2 we explain building the statistical language models using Persian Text Corpus. Section 3 describes the method of incorporating language models in a continuous speech recognition system. In section 4 some experimental results are shown. And section 5 presents some conclusions and future work to be done.

2. Building N-gram Language Models

Persian Text Corpus that we have used to build the language models contains about 8 million words annotated with POS tags. The texts collected in this corpus have been gathered from newspapers, journals, books and etc. The tag set of Persian Text Corpus includes 882 POS tags.

Using this corpus, we have built word-based monogram and POS-based bigram and trigram language models (P. A. Heeman, 1998). Because the Persian Text Corpus doesn't include enough words, the word-based bigram and trigram models will be very sparse. So we preferred to use POS-based bigram and trigram models.

We faced two problems in building the models using Persian Text Corpus. The first problem was orthographic inconsistency of Persian Text Corpus. One of the issues of this problem rises from the fact that Persian writing system allows certain morphemes to appear either as bound to the host or as free affixes – free affixes could be separated by a final form character or with an intervening space. The three possible cases are illustrated for the plural suffix "***hɒ***" (ها) and the imperfective prefix "***mi***" (می):

attached	**final form**	**intervening space**
کتابها (books)	کتاب‌ها	کتاب ها
ketɒbhɒ	***ketɒb~hɒ***	***ketɒb~ hɒ***
میروند (they are going)	می‌روند	می روند
miravand	***mi~ravand***	***mi~ ravand***

In these examples, the tilde (~) is used to indicate the final form marker which is represented as the control character \u200C in Unicode -also known as the zero-width non-joiner (K. Megerdoomian, 2004). All of these surface forms are found in "Persian Text Corpus". For solving this problem we changed all these forms to attached form.

Another issue rises from the use of Arabic script in Persian writing which makes some words have different orthographic realizations. For example three possible forms for word "***masʔulijat***" (responsibility) are:

مسئولیت مسؤولیت مسوولیت

Thus we had to unify these different forms.

The second problem was the multiplicity of POS tags in Persian Text Corpus. As described above, the tag set contains 882 POS tags while many of them contained detailed information about the words. For example, in Persian language each verb can have six different inflectional forms in each tense. So, instead of having just one POS tag for verbs in corpus, there were many different POS tags depending on the tense and the person of the verbs. As an example for this case, the six different forms and their POS tags for the infinitive "***raftan***" (to go) in present tense are shown in table 1.

verb	meaning	POS tag
می روم	I go	V_PRS_POS_1
می روی	you go	V_PRS_POS_2
می رود	he/she goes	V_PRS_POS_3
می رویم	we go	V_PRS_POS_4
می روید	you go	V_PRS_POS_5
می روند	they go	V_PRS_POS_6

Table 1. Six different inflectional forms of verb "go" in present tense in Persian language

The same problem existed for other POS tags like adjectives, nouns, prepositions and etc. So a few numbers of these kinds of POS tags have been used frequently but many of them have been used rarely in texts of corpus. As a solution we decreased the number of POS tags by clustering them manually according to their syntactical similarity. Thus, rare POS tags were classified in larger POS categories. For example, all the POS tags which presented in table 1 and all the negative forms of these tags (e.g. V_PRS_NEG_1) were classified in a larger POS category named "V_PRS".

Beside this redundancy, some of POS tags not only have been rarely used but also were not syntactically significant. So we have used IGNORE tag instead of all unimportant POS tags.

Also a NULL tag has been defined to mark the beginning of any sentences. After considering all above conditions, the size of tag set was reduced to 166 POS tags.

When these problems were solved, we obtained some statistics we need for building language models. These statistics are:

1- The number of times that each word occurs in the corpus (monogram statistics of words).
2- The number of times that each POS tag occurs in the corpus (monogram statistics of POS tags).
3- The number of times that each couple of POS tags occurs in the corpus (bigram statistics of POS tags).
4- The number of times that each triple of POS tags occurs in the corpus (trigram statistics of POS tags).
5- The number of each POS tag that is assigned to each word in the corpus (lexical generation statistics).

As an example for the fifth case of the above statistics, the word "***Zibɒ***" (beautiful) can be considered as either a simple adjective or a proper noun in Persian. Thus we extracted the statistics of each POS tag for this word separately.

Considering the extracted statistics, the monogram probability of each word can be computed as follow:

$$P(w_i) = \frac{N_{monogram}(w_i)}{N_{total}} \tag{2}$$

Where $N_{monogram}(w_i)$ is the monogram statistic for word w_i and N_{total} is the total number of words in the corpus.

The POS-based bigram and trigram probabilities are computed by equation (3) and (4) respectively:

$$P(T_i \mid T_j) = \frac{N_{bigram}(T_j T_i)}{\sum_k N_{bigram}(T_j T_k)} = \frac{N_{bigram}(T_j T_i)}{N_{monogram}(T_j)} \tag{3}$$

$$P(T_i \mid T_k T_j) = \frac{N_{trigram}(T_k T_j T_i)}{\sum_l N_{trigram}(T_k T_j T_l)} = \frac{N_{trigram}(T_k T_j T_i)}{N_{bigram}(T_k T_j)} \tag{4}$$

Where $N_{bigram}(T_j T_i)$ is the number of times that POS tag T_i occurs after POS tag T_j and $N_{trigram}(T_k T_j T_i)$ is the number of times that POS tag T_i occurs after the POS pair $T_k T_j$ in the corpus.

In order to use POS-based bigram and trigram language models we need to have lexical generation probabilities which are computed as follows:

$$P(w_i \mid T_j) = \frac{N(w_i, T_j)}{N_{monogram}(T_j)} \tag{5}$$

Where $N(w_i, T_j)$ is the lexical generation statistics discussed above and $N_{monogram}(T_j)$ is the monogram statistic of POS tag T_j.

Generally the bigram and trigram models have so many zero probabilities because of the sparseness of the data. Thus we smoothed the models with Katz smoothing method (S.M. Katz, 1987).

Our lexicon has about 1090 words which includes lexical generation statistics for each word. For example, the word "***Zibɒ***" (beautiful) has been presented in lexicon as follows:

Zibɒ	ADJ_SIM	420	N_SING_PR	30

In the next section we will discuss the role of statistical language models in continuous speech recognition systems.

3. Incorporating Language Models in Speech Recognition Systems

In general, in speech recognition systems, the language model score can be combined with acoustic model score through two methods: "during search" and "at the end of search" (M. P. Harper et al., 1994). In this paper we have used "during search" method.

In the search process, the sequence of symbols generated by the acoustic component is compared with the set of words present in the lexicon as to produce the optimal sequence of words that will compose the system's final output. In "during search" method when search process recognizes a new word within expanding the different hypothesis, the new hypothesis score is computed via multiplication of following three terms: the n-gram score of new word, the acoustic model score of new word and current hypothesis score. If S_n is the current hypothesis score after recognizing the word W_n and also if W_{n+1} is the next recognized word after expanding the hypothesis, then the new hypothesis score will be:

$$S_{n+1} = S_n . S_{AM}(w_{n+1}) . S_{LM}(w_{n+1})^{LMW} \qquad (6)$$

Where $S_{AM}(w_{n+1})$ is the acoustic model score for word w_{n+1} and $S_{LM}(w_{n+1})$ is its language model score. Because of the difference between scales of $S_{AM}(w_{n+1})$ and $S_{LM}(w_{n+1})$, a weight parameter (LMW) is usually applied to language model score. In general, evaluating equation (6) will lead to problematically small values, so we use the logarithm of probabilities instead of the main probabilities as follow:

$$\begin{aligned}\log S_{n+1} &= \log S_n + \log S_{AM}(w_{n+1}) \\ &+ LMW . \log S_{LM}(w_{n+1})\end{aligned} \qquad (7)$$

In equation (7) for word-based monogram language model $S_{LM}(w_{n+1})$ can be computed by:

$$S_{monogram}(w_{n+1}) = P(w_{n+1}) = \frac{N_{monogram}(w_{n+1})}{N_{total}} \qquad (8)$$

For POS-based bigram and trigram language models, after recognizing each word, the system searches the most probable POS for it and then scores the hypothesis according to that POS.

The most probable POS for word w_{n+1} based on bigram and trigram language models is computed respectively as follow:

$$T_{n+1} = \arg\max_i \left[P(T_i \mid T_n).P(w_{n+1} \mid T_i)\right] \qquad (9)$$

$$T_{n+1} = \arg\max_i \left[P(T_i \mid T_{n-1}T_n).P(w_{n+1} \mid T_i)\right] \qquad (10)$$

Where T_n and T_{n+1} are the most probable POSs for the words w_n and w_{n+1}.

According to equations (9) and (10) for POS-based bigram and trigram language models, $S_{LM}(w_{n+1})$ in equation (7) will be replaced respectively by:

$$S_{bigram}(w_{n+1}) = \max_i \left[P(T_i \mid T_n).P(w_{n+1} \mid T_i)\right] \qquad (11)$$

$$S_{trigram}(w_{n+1}) = \max_i \left[P(T_i \mid T_{n-1}T_n).P(w_{n+1} \mid T_i)\right] \qquad (12)$$

The hypotheses scores are computed by the above equations and finally when the hypotheses completed, the hypothesis with maximum score will be the system's final output. In fact in this method the n-gram language model guides the search process to find the most probable sequence of words.

4. Experimental Results

To evaluate our statistical language models, we used SHARIF speech recognition system (B. Babaali & H. Sameti, 2004), which is a Persian speaker independent continuous speech recognition system. This system performs modeling of monophones using Hidden Markov Model (HMM) and utilizes the word search algorithm described in (S. Ortmanns et al., 1998) for word recognition. In this algorithm, while recognizing the phonemes, the lexicon tree is also searched in order to find the best word sequence according to the phoneme sequence. As described above, the size of vocabulary is about 1090 words.

To run experiments the HMMs were trained for each 30 phonemes of Persian language using 5940 sentences (about 5 hours of read speech) of FARSDAT speech database (M.Bijankhan et al., 1994). We performed the experiments on 140 sentences of FARSDAT database which don't overlap with the training data. As discussed above, because of the difference between scales of acoustic model score and language model score, a weight parameter is usually applied to language model score. In order to choose an optimal weight for language model, we have used our language models with different weights.

In table 2 we present the word error rates (WER) obtained when using different language models with different weights within the speech recognition system. The zero weight is for the case that no language model was used.

LMW	WER Monogram Model [%]	WER Bigram Model [%]	WER Trigram Model [%]
0	34.00	34.00	34.00
1	28.52	26.78	27.82
2	26.44	24.68	26.66
3	26.39	25.03	25.96
4	26.31	25.03	27.36

Table 2. WER obtained by the speech recognition system using different language models and different weights.

The results show that a considerable reduction in word error rates has been achieved by using language models. The maximum reduction in word error rates for monogram, bigram and trigram language models is about 22.6% (LMW=4), 27.4% (LMW=2) and 23.6% (LMW=3) respectively.

Unexpectedly, the trigram language model has reduced word error rate less than bigram language model.

5. Conclusions

This paper reported our work on developing language models for a Persian continuous speech recognition system. The results showed the effect of language models in guiding the search process which increases the accuracy of speech recognition system. Also the importance of language model weight in reduction of the word error rate was demonstrated through the experiments.

Although we expected that the performance of POS-based trigram language model would be better than other models but the results showed that POS-based bigram language model have the better performance than POS-based trigram language model. We think that the reason is impropriety of POS-based n-gram language models for prediction the next word.

The next stage in this research will be clustering of Persian words by common word clustering methods (P. Brown et al., 1992) and building the class-based n-gram language models.

6. References

Babaali, B. & Sameti, H. (2004). The Sharif Speaker-Independent Large Vocabulary Speech Recognition System. *In proceedings of 2nd Workshop on Information Technology & Its Disciplines*. Kish Island, Iran.

Bijankhan M. et al. (1994). FARSDAT-The Speech Database of Farsi Spoken Language. *In proceedings of the 5th Australian International Conference on Speech Science and Technology*. (vol. 2).

Brown, P. & Della Pietra, V. & deSouza, P. & Lai, J. & Mercer, R.L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 467—479.

Harper, M.P. & Jamieson, L.H. & Mitchell, C.D. & Ying G. & Potisuk, S. & Srinivasan, P.N. & Chen R. & Zoltowski, C.B. & McPheters, L. L. & Pellom, B. & Helzerman, R.A. (1994). Integrating Language Models with Speech Recognition. *In proceedings of AAAI-94 Workshop on the Integration of Natural Language and Speech Processing*, (pp. 139-146).

Heeman, P.A. (1998). POS tagging versus Classes in Language Modeling. *In proceedings of 6th Workshop on Very Large Corpora*, (pp. 179-187).

Huang, X. & Alleva, F. & Hon, H. & Hwang, M. & Lee, K. and Rosenfield, R. (1993). The SPHINX-II Speech Recognition System: An Overview. *Computer Speech and Language*. (vol. 2, pp. 137-148).

Katz, S.M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transaction on Acoustics, Speech and Signal Processing*, (ASSP-35 (3), pp 400-401).

Megerdoomian, K. (2004).Finite-State Morphological Analysis of Persian. *In proceeding of 20th International Conference on Computational Linguistics*, (pp 35-41).

Ortmanns, S. & Eiden, A. & Ney, H. (1998). Improved Lexical Tree Search for Large Vocabulary Speech Recognition. *In proceedings of IEEE International Conference on Acoustics, Speech and Signal Proc.*

Rabiner, L. & Juang, B.H. (1993). Fundamentals of Speech Recognition. New Jersey: Prentice Hall.

Young, S.J. & Jansen, J. & Odell, J.J. & D. Ollason, D. & Woodland, P.C. (1995). The HTK Hidden Markov Model Toolkit Book.