

Affix-Augmented Stem-Based Language Model for Persian

Hesham FAILI

Department of ECE, University of Tehran.

Tehran, Iran

hfaili@chamran.ut.ac.ir

Hadi RAVANBAKSH

Department of ECE, University of Tehran.

Theran, Iran

h.ravanbakhsh@ece.ut.ac.ir

Abstract:

Language modeling is used in many NLP applications like machine translation, POS tagging, speech recognition and information retrieval. It assigns a probability to a sequence of words. This task becomes a challenging problem for high inflectional languages. In this paper we investigate standard statistical language models on the Persian as an inflectional language. We propose two variations of morphological language models that rely on a morphological analyzer to manipulate the dataset before modeling. Then we discuss shortcoming of these models, and introduce a novel approach that exploits the structure of the language and produces more accurate analysis with less amount of training data. Experimental results are encouraging especially when we use n-gram models with small training dataset.

Keywords:

Tracking; language model; n-gram; morphological; Persian

1. Introduction

A standard statistical language model (LM) calculates the probability of a word sequence $W = w_1, w_2, \dots, w_n$ as a product of the conditional probabilities of each word w_i given its history, which is typically approximated by the one or two most recent words. A commonly used trigram model can be applied to a word based language model to achieve this goal.

However this approach is not effective for inflectional languages. These languages with a relatively free word-order (e.g. Russian and Persian) necessitate a different and more sophisticated word model [11]. The problem with direct application of n-gram methods is that as a result of free word order, many word contexts are infrequent and estimated conditional probabilities are not reliable. The problem escalates with morphologically rich languages such as Turkish, Russian and Arabic, which have a high vocabulary growth rate. This brings about even higher model perplexity and large number of out of vocabulary words [14]. Also, language modeling becomes more problematic when it related to languages with low resource availability. The lack of training corpus in these kinds of languages leads the model to be sparse.

To overcome these problems, morphological approaches for language modeling have been proposed. Early approaches on morphological based language modeling presented in [4] introduce class-based models, using part-of-speech of words as class. Whittaker's solution to this problem for Russian language is to extract more information from the word itself rather than its position in the sentence. In a well-defined and unambiguous decomposition he derives parts of each word which he calls particles and then applies the n-gram model after such decomposition [15]. A new type of language model called factored language model [1], which uses morphological information in a novel back-off procedure. This model used in [7] for Arabic and compared to particle model, class based model and single-stream model, where sequences of stems, morph tags, etc. are considered individually.

In Persian, we observe the same phenomenon of high perplexity, resulting from inflectional property. Especially, this property is more severe in the case of verbs, since they are the most inflectional words of Persian. This could result in higher perplexity due to critical role of the verbs in all sentences. In this paper we introduce a new approach to language model construction called "affix augmented stem based language model" to overcome the problem.

The remainder of paper is organized as follows. The next section introduces n-gram model as the base for language model. In section 3, we introduce two well-known morphological language models including stem-based and morph-based approaches, that implemented in this article as well as affix augmented stem based language model. Section 4 presents the experimental results comparing these models and finally we conclude with a summary of future work.

2. Language Model

A language model determines the probability $P(W_1^n)$ of a word sequence W_1^n . This probability is decomposed as follows:

$$P(W_1^n) = \prod_{i=1}^n P(W^i | W_1^{i-1}) \quad (1)$$

We refer to $P(W^i | W_1^{i-1})$ as conditional probability of W^i in this paper.

n-gram is the most popular approach in language

modeling in which the model represents the conditional probability of a word given $n-1$ previous words. By increasing n the accuracy of the model will be enhanced due to capturing more information [9]. However for small training sets the result would not be reliable as the number of samples based on which we are estimating these probabilities is not large enough. Even with reasonably small n , many n -gram patterns may not appear in the data set. This is known as data sparseness problem. Smoothing methods have been used to resolve this inadequacy of data [8].

We have two approaches for the evaluation of language models: *Extrinsic* approach, which evaluates the model in some applications using model as a component; and *Intrinsic* evaluation, which evaluates model, just by using the model itself, with a test-data. The later method typically has been applied using *Perplexity* as accuracy measure. This measure is related to the concept of entropy which measures the average uncertainty in the value of a random variable. Capturing more knowledge or structure by a model would lower the uncertainty or entropy. Therefore; we consider a model with the lower entropy as a better model. The perplexity (PP) can be defined as the average number of possible words following any string of $n-1$ words in a large corpus based on n -gram language model [12]:

$$PP = 2^{\frac{1}{n} \log_2(P(W_1^n))} \quad (2)$$

Where n is the size of the test corpus and W is the test data word sequence.

We use open vocabulary model which assign <unk> label to words, which are not in dictionary in both train and test phase.

3. Morphological Language Models

In a simple n -gram model, the probability of a word only depends on $n-1$ previous words. In fact $P(W^i|W_1^{i-1})$ is replaced with $P(W^i|W_{i-n+1}^{i-1})$ in (1). We calculate these probabilities in the training phase; then we use them to measure PP in the evaluation phase. We refer to this model as word based language model (WB-LM) as the base model. The problem of this model in inflectional language is its high rate of OOV, which is the result of highly inflectional words. Consider a stem of observed word; WB-LM can not recognize the stem with other affixes unavailable in the training data. A first thought is to use a larger training set. However, since such dataset should contain all possible combinations of each stem and all possible affixes, this is not a practical solution.

One realistic approach is to analyze the dataset, morphologically.

Morphological-Based language models are used for

highly inflectional languages in order to address this problem. In the following we represent three such language models for Persian language.

Stem-Based language model (SB-LM) removes the affixes both in the training phase and in the evaluation phase and uses the new train and test data as an input of the WB-LM. It helps to deal with many OOV words in WB-LM. However this is just a quick fix and does not completely solve the problem for highly inflectional languages such as Arabic. This is because in these languages, affixes usually play an important role in meaning and grammar.

The lost, incurred by ignoring affixes in SB-LM resulted in more interest in affixes. The final result is Morphological based language models (MB-LM). In this kind of model, we decompose all the words in the dataset into their morphemes and each of them is considered as a complete word. Then by using this new dataset a WB-LM similar to the SB-LM is learned. This method is similar to the Morpheme-based language model used in [9], but here we assumed particles are only one prefix, one stem and one suffix. This language model decreases the rate of OOV, either. Also, affixes as separate words are so frequent that all would occur in a typical dataset and it is very likely for them to get a high conditional probability and this language model retrieves a better PP.

MB-LM has its own problems. First, it increases the number of words, and in order to capture the structure of the language, we need to use higher n -grams. This results in more memory requirement. Furthermore, this method exaggerates the grammatical importance of affixes in the language of our interest. For example in Persian, the prefix of a word often does not depend on the previous gram directly, which might be the suffix of previous word. As a consequence the model is encumbered with such unnecessary or even misleading information.

3.1. Affix-Augmented Stem-Based Language Model (AASB-LM)

In this section, a new approach that combines the benefits of SB-LM and MB-LM is presented. We name it to *affix augmented stem based language model (AASB)*.

The key idea is that in Persian, affixes are mainly coupled with their stem and other affixes of the same stem. Also these affixes contain little information about other words. We observed that affixes depend on some surface lexical information (SLI) of the stem. The mentioned general idea however was too simplistic and overlooked the following details.

First, dependencies of affixes to SLI are not generally true when we use POS tags as SLI for regular POS tags in Persian. For this assumption to be held, we need to enrich each POS with some augmented information. For example, imperative verbs ending with “و” /oo/ like “گو” /goo/ use “ید” /yæd/ instead of “د”

/æd/ as suffix. Therefore we can have two SLI for imperative verb; the one with “د” /æd/ as its acceptable suffix and the one with “يد” /yæd/.

Second, an affix should be selected so that SLI of the stem remains the same as SLI of the word, otherwise, the role of the word in the sentence may be changed and as a consequence, the conditional probabilities of two partially independent words will mix together. For example the n -gram of “چوب” /choob/ “wood” and “چوبی” / choobi / “wooden” are not the same, because the first is a noun, and the second is an adjective. Therefore to prevent such problems we do not consider “ی” /e/ as a suffix of a noun in the morphology-analyzing phase, but put words in format of noun+“ی” /e/ to dictionary as adjectives.

Third, each word should get only one SLI. This can be easily obtained by combining SLIs that a word belong to, and make a new SLI. For example word “دو” / do (dav) / run, is belong to past verb (<PV> SLI) and imperative verb (<IV> SLI); therefore we should define a new SLI as <PV-IV> and every word like “دو” do not belong to <PV> or <IV> but <PV-IV> in new set of SLIs. Furthermore, potential affixes for <PV-IV> are potential affixes for <PV> or <IV>.

In the proposed AASB-LM, in addition to the n -gram conditional probability of stems, two conditional probabilities for affixes are also trained. The conditional probabilities are as follow:

$P(\langle \text{prefix} \rangle | \langle \text{SLI} \rangle)$ which is the probability of <prefix> happening before the <SLI>, so that <prefix> is a prefix for the surface lexical information <SLI>.

$P(\langle \text{suffix} \rangle | \langle \text{SLI} \rangle, \langle \text{prefix} \rangle)$ which is the probability of <suffix> happening after the <SLI> which has <prefix> as its prefix. Also <prefix> can be <null>.

After training these probabilities, model is ready to compute probability of a given sentence. This model takes different methods to compute conditional probabilities for prefixes, suffixes and stems in order to gain a lower PP. Therefore, we need to decompose each word to morphemes in the dataset. After the decomposition, $P(W)$ can be computed by the following generative story:

First, for any stem w_i , conditional probability is calculated in a similar fashion to SB-LM. That is:

$$P(W^i | W_1^{i-1}) = P(W^i | W_{i-n+1}^{i-1}) \quad (3)$$

Where $\{w_j\} (i-n-1 < j < i)$ are $n-1$ previous stems.

Then we calculate the conditional probability for affixes as follow:

For suffixes:

$$P(\langle \text{suffix} \rangle | W_1^{i-1}) = P(\langle \text{suffix} \rangle | \langle \text{SLI}_{stem} \rangle, \langle \text{prefix} \rangle) \quad (4)$$

This conditional probability can be computed easily using MLE method in the training phase and it can be used directly. Suffix’s dependency on SLI of stem and prefix is suggested by formula (4). <prefix> is <null> if the word has no prefix.

Computing prefix conditional probability is not as easy as that of suffix. The main assumption is that an affix depends on the SLI of the stem. The point is that the conditional probability is calculated based on its previous $n-1$ grams, but as stem comes after prefix and we can not use the SLI of the stem. In order to overcome this, we decompose the problem in this way:

$$P(\langle \text{prefix} \rangle | W_1^{i-1}) = P(\langle \text{prefix} \rangle | W_{i-n+1}^{i-1}) = \sum_{SLI \in VSLIs} (P(\langle \text{prefix} \rangle | SLI, W_{i-n+1}^{i-1}) P(SLI | W_{i-n+1}^{i-1})) \quad (5)$$

Where $VSLIs$ are all valid SLIs that can accept <prefix> as their prefix.

In fact we consider each SLI with <prefix> as its potential prefix and then calculate the probability of the SLI as the surface lexical information of the next stem that has not appeared yet. Therefore, first we should sum all the conditional probabilities of stems with the SLI as their surface lexical information, giving $n-1$ recent stems:

$$P(\langle \text{SLI} \rangle | W_{i-n+1}^{i-1}) = \sum_{stem \in \langle \text{SLI} \rangle} P(stem | W_{i-n+1}^{i-1}) \quad (6)$$

This is the probability of the next stem with <SLI> as its SLI. Moreover, prefix is conditionally independent of previous $n-1$ gram, therefore:

$$P(\langle \text{prefix} \rangle | SLI, W_{i-n+1}^{i-1}) = P(\langle \text{prefix} \rangle | SLI) \quad (7)$$

In order to prevent high running time complexity in test phase, we should compute conditional probabilities of SLIs and put them in model in training phase.

Here we obtained a language model, uses different strategies to calculate conditional probabilities for prefix stem and suffix, given previous words.

AASB-LM has its own problem, which is the fundamental assumption used in equation (3), (4) and (7). In some application of language model like grammar checker, this could result in bad language modeling and give lower PP to wrong grammar structures while MB-LM gives better results. For example, pattern “+ ها <adjective> + ها + <adjective>” (Where “ها” /hΛ/ is suffix for adjectives and nouns.) is not a valid structure and AASB-LM can give high probability to this phrase; however these kind of situations hardly happen.

3.2. An Illustrative Example

AASB Here we present an example to calculate the

PP for a Persian sentence, using 3-gram AASB-LM. The input sentence is:

“من با دوستانم به مدرسه می روم”

/mæn bʌ doostʌnæm be mædrese mi rævæm/

“I am going to school with my friends.”

In AASB-LM, the primary sentence is decomposed to the sentence below:

“من با دوست/انم به مدرسه می/رو/م”

While “انم” /ʌnæm/ and “م” /æm/ are the suffixes and “می” /mi/ is the prefix.

The conditional probabilities for stems are:

$$\begin{aligned} P(\text{من} \mid \langle \text{begin} \rangle) &= P(\text{من} \mid \langle \text{begin} \rangle) \\ P(\text{با} \mid W_1^1) &= P(\text{با} \mid \langle \text{begin} \rangle, \text{من}) \\ P(\text{دوست} \mid W_1^2) &= P(\text{با}, \text{دوست} \mid \text{من}) \\ P(\text{به} \mid W_1^4) &= P(\text{دوست}, \text{با} \mid \text{به}) \\ P(\text{مدرسه} \mid W_1^5) &= P(\text{به}, \text{دوست} \mid \text{مدرسه}) \\ P(\text{رو} \mid W_1^7) &= P(\text{مدرسه}, \text{به} \mid \text{رو}) \\ P(\langle \text{end} \rangle \mid W_1^9) &= P(\text{رو}, \text{مدرسه} \mid \langle \text{end} \rangle) \end{aligned}$$

The conditional probabilities for suffix are:

$$P(\text{انم} \mid W_1^3) = P(\langle \text{Countable-Noun} \rangle, \langle \text{null} \rangle \mid \text{انم})$$

$$P(\text{می} \mid W_1^8) = P(\langle \text{ImperativeVerb-NotEndIn} \rangle \mid \text{می})$$

While $\langle \text{Countable-Noun} \rangle$ and $\langle \text{ImperativeVerb-NotEndIn} \rangle$ are SLIs of “دوست” and “رو” respectively.

Here we assume $\langle \text{Past-Verb} \rangle$ ($\langle \text{PV} \rangle$), $\langle \text{Imperative-Verb-EndIn} \rangle$ ($\langle \text{IVU} \rangle$) and $\langle \text{Imperative-Verb-NotEndIn} \rangle$ ($\langle \text{IVNU} \rangle$) are the SLIs that can accept “می” as their prefix (VLSIs = $\{\langle \text{PV} \rangle, \langle \text{IVU} \rangle, \langle \text{IVNU} \rangle\}$).

Probability of prefix “می” /mi/ by equation (5) and (7) is equal to:

$$\begin{aligned} P(\text{می} \mid W_1^6) &= \\ P(\text{می} \mid \langle \text{PV} \rangle) &\times P(\langle \text{PV} \rangle \mid \text{مدرسه}, \text{به}) \\ + P(\text{می} \mid \langle \text{IVU} \rangle) &\times P(\langle \text{IVU} \rangle \mid \text{مدرسه}, \text{به}) \\ + P(\text{می} \mid \langle \text{IVNU} \rangle) &\times P(\langle \text{IVNU} \rangle \mid \text{مدرسه}, \text{به}) \end{aligned}$$

Now we can calculate unavailable probabilities according to equation (6), summing over all available verbs. For example considering $\langle \text{PV} \rangle$ SLI:

$$\begin{aligned} P(\langle \text{PV} \rangle \mid \text{مدرسه}, \text{به}) &= \\ P(\text{رفت} \mid \text{مدرسه}, \text{به}) &+ P(\text{دو} \mid \text{مدرسه}, \text{به}) \\ + P(\text{گفت} \mid \text{مدرسه}, \text{به}) &+ P(\text{رو} \mid \text{مدرسه}, \text{به}) + \dots \end{aligned}$$

While “رفت” /raft/ went, “دوید” /dæved/ ran, “گفت” /goft/ said and “خواند” /khʌnd/ read, are words with $\langle \text{PV} \rangle$ as their SLIs.

4. Experiments and Analysis

The data used in the experiments is from Hamshahri² newspaper archive. The text contains 400K

sentences, which we divided it into two datasets for training (350K sentences) and testing (50K sentences) purpose. We evaluated four language models, discussed earlier. We also performed another experiment with small training set of 70K sentences.

For all experiments, we used training data to calculate n-gram probabilities using SRILM toolkit [13]. This toolkit uses Good Turing smoothing [6] to overcome the problem of data sparseness. In the evaluation phase, we ignore all words with $\langle \text{unk} \rangle$ labels—similar to SRILM in its evaluation phase.

Both training and test data were manipulated to fit different experiments. For this, we used a morphology analyzer that can divide each word into prefix, stem and suffix. Morphology analyzer uses a dictionary of stems with their SLIs and a list of all affixes. It is mentioned for each affix potential SLIs for stems they can be attached to. Also, the POS tag set is borrowed from [10] and adapted to our experiments by annotating other information to obtain SLIs (see Section 3).

Morphological LMs (SB-LM, MB-LM, AASB-LM), decrease the OOV rate, and therefore we cannot properly compare their PP against that of WB-LM. Table 1 shows our results for both PP and OOV. This experiment confirms that the perplexity of WB-LM is much higher due to the different OOV. Therefore the morphology-analyzer should be the main factor in reducing PP.

Table 1. Perplexity and OOV Rate of Different Language Models in 3-gram with Small Train Date

Language Model	Perplexity	OOV rate
WB-LM	1508	15%
SB-LM	422	12%
MB-LM	367	12%
AASB-LM	342	12%

Figure 1 and Figure 2 show the perplexity of the mentioned morphological language models for different value of n , in n-gram model, trained on small and large dataset respectively.

One could notice that construction of better stem based models (SB-LM and AASB-LM), is possible with smaller training sets compared to MB-LM. This is, due to the methods to assign reliable conditional probabilities. It is not necessary for all possible combinations of a stem and its affixes to appear in the training dataset. Stem-based models therefore need less data to learn a good model and perform better on small training data sets, however they are also competitive with MB-LM when using large training sets.

² Hamshahri is one of the most popular daily newspapers in Iran that has been publishing for more than 20 years. Hamshahri2 corpus is a Persian test collection that consists of

1.4 GB of news texts from this newspaper since 1996 to 2007.

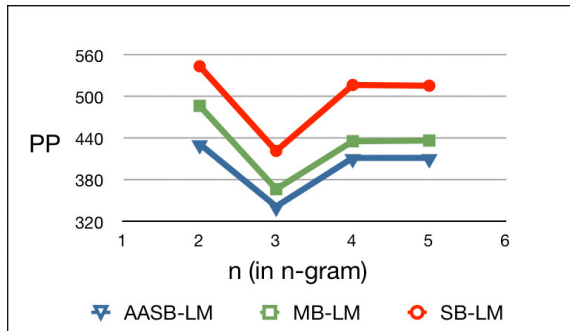


Figure 1. Perplexity Respect to Different n in n-gram Model (Small Trainset)

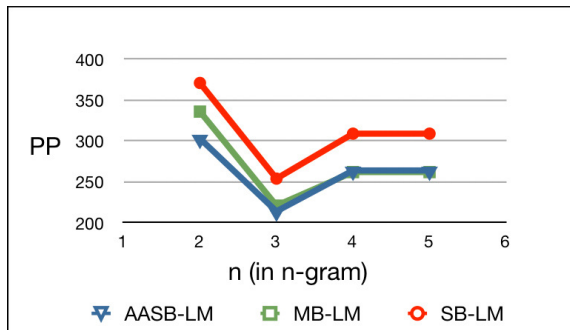


Figure 2. Perplexity Respect to Different n in n-gram Model (Large Trainset)

The unintuitive increase in PP for 4-gram (Fig. 1) is because of data sparseness and the problem cannot be resolved by increasing the size of training set as shown in Fig. 2. This anomaly (of higher perplexity with larger n) has also been observed in [9] for Arabic. This phenomenon appears in lower degrees for MB-LM (Fig. 2), because in these models, affixes are among previous grams, which occur more frequently and thus conditional probabilities given 3 previous grams, are usually higher.

5. Conclusion and Future Work

We showed that SB-LM and MB-LM are models of choice for Persian, producing more accurate models compare to WB-LM. We introduced a new language model, named AASB-LM, and demonstrated its better accuracy compared to other language models. We have shown this model even reduces the perplexity 7% compared to MB-LM, as the best available model for inflectional languages.

We believe further research can improve our morphological analyzer and language model so that words in form of prefix*-stem-suffix* (where * denotes zero or more occurrence of morphemes.) can be detected and each morpheme becomes a separate word. More extensive experiments on LMs, can assess the robustness of different methods to wrong data for evaluation.

Extrinsic evaluation is another direction in which we can evaluate the new model in applications such as speech recognition. Finally we plan to use our language model for other inflectional languages (especially languages with high rate of prefixes such as Turkey). In our languages of interest, affixes should depend on morphemes of the word to which they refer and are rather independent of other words in order to apply on our language model.

References

- [1] J. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies, pp. 4–6, 2003.
- [2] M. Elbeze and A. Derouault, "A morphological model for large vocabulary speech recognition," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 577–580, 1990.
- [3] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 35, pp. 400–401, 1987.
- [4] K. Kirchhoff, D. Vergyri, J. Bilmes, D. Duh, and A. Stolcke, "Morphology-based language modeling for Arabic speech recognition," In Proceedings of International Conference on Spoken Language Processing, vol. 3, pp. 2245–2248, 2006.
- [5] X. Liu and W. B. Croft, "Language modeling for information retrieval," Annual Review of Information Science and Technology, vol. 39, pp. 3–31, 2005.
- [6] K. Meftouh, K. Smaili, and M. Laskri, "Statistical modeling of Arabic language," Proceedings of 9e Journées internationales d'Analyses statistique des Données Textuelles, pp. 12–14, 2008.
- [7] K. Megerdooomian, "Persian computational morphology: a unification-based approach," Conference on Intelligent Text Processing and Computational Linguistics, pp. 135–149, 2000.
- [8] I. Oparin and A. Talanov, "Stem-based approach to pronunciation vocabulary construction and language modeling of Russian," Proceeding of the 10th. International Conference on Speech and Computer, pp. 575–578, 2005.
- [9] L. Rabiner and B. Juang, Fundamental of Speech Recognition, 1rd ed., Prentice Hall, New Jersey, 1993.
- [10] A. Stolcke, "SRILM - an extensible language modeling toolkit," Proceedings of the International Conference on Spoken Language Processing, pp. 901–904, 2002.
- [11] D. Vergyri, K. Kirchhoff, K. Duh, and A. Stolcke, "Morphology-based language modeling for Arabic speech recognition," Computer Speech & Language, vol. 20, pp. 589–608, 2004.
- [12] E. W. D. Whittaker, "Statistical language modeling for automatic speech recognition of Russian and English," Ph.D. Thesis, Cambridge University, 2000.