A new model for Persian multi-part words edition based on statistical machine translation

M. Zahedi* and A. Arjomandzadeh

School of Computer Engineering & Information Technology, University of Shahrood, Shahrood, Iran.

Received 07 July 2014; Accepted 04 July 2015
*Corresponding author: zahedi@shahroodut.ac.ir (M. Zahedi).

Abstract
Multi-part words in English language are hyphenated and hyphen is used to separate different parts. Persian language consists of multi-part words as well. Based on Persian morphology, half-space character is needed to separate parts of multi-part words where in many cases people incorrectly use space character instead of half-space character. This common incorrectly use of space leads to some serious issues in Persian text processing and text readability. In order to cope with the issues, this work proposes a new model to correct spacing in multi-part words. The proposed method is based on statistical machine translation paradigm. In machine translation paradigm, text in source language is translated into a text in destination language on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. The proposed method uses statistical machine translation techniques considering unedited multi-part words as a source language and the space-edited multi-part words as a destination language. The results show that the proposed method can edit and improve spacing correction process of Persian multi-part words with a statistically significant accuracy rate.

Keywords: Persian Multi-Part Words, Spacing Rules, Statistical Machine Translation, Parallel Corpora, Hierarchical Phrase-based, Fertility-based IBM Model, Syntax-Based Decoder.

1. Introduction
Persian text consists of words which are made of multiple parts and they are called multi-part words. An important key note in multi-part words is that the parts of multi-part words must be separated while whole multi-part word must be distinguished as an integrated word; To achieve this goal, the parts of multi-part words must be separated by half-space character to keep the integrity of whole multi-part word. Half-space is a character with zero-width non-jointer length which is actually used to prevent joining the characters of the multi-part words and keep the parts of multi-part word as close as possible. One of the most common problems in Persian text is incorrectly use of spaces between multi-part words which leads to non-integrity of multi-part words and it also leads to incorrect word boundary detection that can be solved by replacing spaces with half-spaces. Based on Persian language spacing rules which specify where space or half-space is needed, half-spaces must be inserted between parts of multi-part words. If space character is used between the parts of multi-part words, the word does not obey standard word form and each part will be incorrectly considered as a separate word such as، "همیشه‌ی یک شمار" and "حامل ضرب". It is important to be noticed that the spell checker algorithms concentrate on the spelling errors which are often caused by operational and cognitive mistakes [1], thus the errors occurring due to the usage of space and half-space in a wrong manner are usually ignored by spell checker algorithms.

Few researchers have worked on editing the spacing in Persian words [2-4]. A toolkit is presented by Shamsfard et al. [2] to detect boundaries of words, phrases and sentences, check and correct the spelling, do morphological analysis and Part-Of-Speech (POS) tagging. The approach finds the stems and affixes of words with Finite State Automaton (FSA) and tags them
focused only on modeling Persian verb morphology. The method detects six morphological features of a given verb and generates a verb form using a FSA. These features consist of several language-specific features such as POS of a given verb, dependency relationships of the verb and POS of subject of the verb. Consequently, unsupervised clustering is used to identify compound verbs with their corresponding morphological features in the training step. In this approach POS taggers are used by a statistical method in order to extract some features and FSA is employed to generate an inflected verb form using these morphological features. Rasooli et al. [4] provide a lexicon which consists of space-separated multi-part words that are mapped to half-space separated multi-part words. The approach identifies all the space-separated multi-part words that can be mapped to half-space separated multi-part words. An expanded lattice version of the sentence including both forms is then decoded with a language model to select the path with the highest probability. This approach relies on a lexicon which consists of all kinds of Persian multi-part words such as verb inflections. Therefore, if the lexicon lacks in multi-part words, the approach cannot edit spaces between the parts of word efficiently. The aforementioned approaches rely on lexicon. So, if the lexicon lacks in multi-part words, the approach cannot edit spacing in multi-part words.

The main issue in POS tagger approach is lexicon that must cover all the variety of the multi-part words in which all the parts of the multi-part words are tagged. On the other hand, the lack of the tagging especially in half-space rule leaves more unedited multi-part words in evaluation step. Moreover, available Persian tagged corpus such as Peykare [5] does not comply with half-space character. In this paper, we propose a different statistical approach which uses a fertility-based IBM Model [6] as word alignment by employing a parallel corpus which is created for the special purpose of Persian multi-part word edition. In the next step, Synchronous Context-Free Grammar (SCFG) for hierarchical phrase-based translation [7] is employed. In decoding step, the extracted grammars and weights assigned to each grammar are employed to decode the word with a syntax-based decoder.

This paper is organized as follows. In section 2, the problems and challenges of Persian text space rules and machine translation theory are reviewed. Section 3 describes fertility-based IBM model and hierarchical phrase-based and utilizes the proposed method in order to edit spacing in Persian text. The next section discusses experimental results and finally the paper ends with conclusion section.

2. Preliminaries

2.1. Spacing issues

In the standard morphology of Persian text, parts of multi-part words should be separated with zero-width non-jointer length character. Therefore, if space character is used in multi-part words, the parts are incorrectly considered as separate words. Space character specifies boundaries of words and half-space character is used for separating the parts in multi-part words.

Based on standard morphology of Persian text, there are two types of spacing between words:
- Spacing between words in a sentence, which is called “space”.
- Spacing between the parts of multi-part words which is called “half-space”. Some words are made up of several parts, but the parts make up a single word which are called multi-part words, such as:

�يرغلبل، پی، حوصله، پایین، تر، جریمه‌ها، رابیت‌های، ازدست‌نمی‌دهد، می‌شود، نصوم نموده‌اند، می‌باشد

Half-space is a character with zero-width non-jointer length which is actually used to prevent joining the parts in multi-part words and keep the parts of multi-part word as close as possible. The terms "می‌شود" and "زبان‌شناسی" are made up of two parts in which half-space maintains word integrity in these multi-part words.

Correct word spacing specifies correct word boundaries which is denoted by spaces in Natural Language Processing (NLP) and clears ambiguity of text. Word boundary detection is considered as an important first step in Persian natural language processing tasks. Half-space character is important in word boundary detection in cases where Persian words are made up of multiple parts.

2.2. Basic theory of statistical machine translation

In Statistical Machine Translation (SMT) theory, every word in source language has many translations and highest probability in corpora (which is defined by (1)) is assigned to the most appropriate translation. Due to Bayes theorem (which is defined by (2)) and since the denominator here is independent of e, finding \( \hat{e} \) is the same as finding e. So, to make the product
Given a vector alignment of a source sentence \( a_i^j \), the fertility of target word \( i \) expresses the number of source words aligned to it \([11]\).

\[
\Phi_i(a_i^j) = \sum_{j=1}^{a_i^j} 1
\]

It omits the dependency on \( a_i'^j \) (and defining \( P(j \mid 0)=1 \)), the probability is expressed as follows.

\[
P(i_j, a_i^j \mid e_i) = P(\Phi_i \mid j) . \prod_{i=1}^{a_i^j} P(\Phi_i \mid e_i) \prod_{j=1}^{a_j} P(i_j \mid a_j)
\]

For each foreign input word \( j \), it factors on the fertility probability \( P(\Phi_i \mid j) \). The factorial \( \Phi_i \) stems from the multiple tableaux for one alignment, if \( \Phi_i > 1 \).

To compute the translation model probability, a fertility-based IBM Model is employed as insertion words (NULL insertion) and dropping of words (words with fertility 0) to edit the multi-part words spacing. Sentence alignment in figure 1 is shorthand for a theoretical stochastic process by which unedited words would be changed into edited words. There are a few sets of decisions to be made. As an example, the word “محمدزاده” is a multi-part word which consists of “محمد” and “ژاده”. So, the space character between the two parts must be edited into half-space character.

The proposed method employs hierarchical phrase-based translation to model half-space in phrases. Hierarchical phrase-based translation is a translation model based on synchronous context-free grammars that models translation as phrase pairs. The translation rules are extracted from parallel aligned sentences \([7]\). On the other hand, hierarchical phrase-based translation employed IBM Model word alignment to extract hierarchical phrase pairs. Therefore, it extracts structure of multi-part words and employs the extracted grammars to edit the multi-part words.

3.2. Proposed method

The general procedure of proposed approach consists of accompanying general methodology of SMT: word alignment, build hierarchical phrase-based model using Synchronous Context-Free Grammar (SCFG), Training phase for weighting extracted features in log-linear model with minimum error rate training and decoding.

In the first phase, words are aligned based on IBM model. In the second phase of the proposed approach hierarchical phrase-based model is employed to extract synchronous context-free grammar. Grammar extraction needs a symbol character to extract linguistic information of space and half-space while space character and half-space character are not considered as symbol characters. In the proposed approach token “*”
and token "&" are chosen to denote space character and half-space character, respectively. Therefore, grammar extraction extracts linguistic information of space character between the distinct words and half-space character between the parts of multi-part words. In the third phase, Log-linear model is trained with MERT. MERT determines weights which denote the importance level of grammars. The proposed approach uses a log-linear model with seven features. To avoid trying to support all the multi-part words in dataset, the structure of multi-part words is trained by the training dataset. To do this, the approach needs linguistic information about space character between the distinct words and half-space character between the parts of multi-part words. The created parallel corpora contain 30000 words which contains various multi-part words with different number of occurrences. A sample of created parallel corpora is presented in table 1. As shown in table 1, the structure of parallel corpora consists of unedited multi-part words in source side and the edited one in the target side in which token "\(\alpha\)" denotes space character and token "\(\&\)" denotes half-space character.

| Table 1. a sample of created parallel corpus. The left side is unedited corpus and the right side is edited corpus. |
|---------------------------------|---------------------------------|
| **unedited corpus**             | **edited corpus**               |
| **Text**                        | **Text**                        |
| یه‌ی گوشش خوردن، مرکز ۱۵ را حکیمی، سرگذاری صحیح، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | یه‌ی گوشش خوردن، مرکز ۱۵ را حکیمی، سرگذاری صحیح، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | یه‌ی گوشش خوردن، مرکز ۱۵ را حکیمی، سرگذاری صحیح، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای |
| گوشش خوردن. مرکز ۱۵ را حکیمی، سرگذاری صحیح. به‌کارگیری چرخه. رابطه. بی‌بی‌کی‌سه. شاره‌ای | گوشش خوردن. مرکز ۱۵ را حکیمی، سرگذاری صحیح. به‌کارگیری چرخه. رابطه. بی‌بی‌کی‌سه. شاره‌ای |
| نشانه‌ی صحیح. به‌کارگیری چرخه. رابطه. بی‌بی‌کی‌سه. شاره‌ای | نشانه‌ی صحیح. به‌کارگیری چرخه. رابطه. بی‌بی‌کی‌سه. شاره‌ای |
| نشانه‌ی صحیح. به‌کارگیری چرخه. رابطه. بی‌بی‌کی‌سه. شاره‌ای | نشانه‌ی صحیح. به‌کارگیری چرخه. رابطه. بی‌بی‌کی‌سه. شاره‌ای |
| نشانه‌ی صحیح. به‌کارگیری چرخه. رابطه. بی‌بی‌کی‌سه. شاره‌ای | نشانه‌ی صحیح. به‌کارگیری چرخه. رابطه. بی‌بی‌کی‌سه. شاره‌ای |
| نشانه‌ی صحیح. به‌کارگیری چرخه. رابطه. بی‌بی‌کی‌سه. شاره‌ای | نشانه‌ی صحیح. به‌کارگیری چرخه. رابطه. بی‌بی‌کی‌سه. شاره‌ای |
| مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای |
| مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای |
| مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای |
| مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای |
| مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای |
| مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای |
| مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای |
| مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای |
| مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای | مثالی از کاتالیزور، به‌کارگیری چرخه، رابطه، بی‌بی‌کی‌سه، شاره‌ای |

Figure 2 shows an overview of the proposed method. In the first phase, words are aligned based on IBM model. The standard way of aligning word is the method implemented in GIZA++ [12, 13]; In the next phase, Thrax grammar extractor is used to extract SCFGs with the aid of Hadoop method that is applicable to large datasets [14]. It also supports extraction of both Hiero [7] and SAMT grammars [15] with extraction heuristics.

The last phase includes training and testing. Z-MERT [16] is used in training step to extract K-best candidate translation. Log-linear employed Minimum Error Rate Training (MERT) [17] method with Z-MERT toolkit in the training step to tune parameters. Seven parameters are tuned in this step:

- N-gram language model \(P_{LM}(t)\) parameter, lexical translation model \(P_c(\gamma|\alpha)\) parameter and \(P_c(\alpha|\gamma)\) parameter, rule translation model \(P(\gamma|\alpha)\) parameter and \(P(\alpha|\gamma)\) parameter, word penalty parameter and the arity of word parameter. Regarding rules of the form \(X \rightarrow \gamma, \alpha, \cdots, \omega\) in hierarchical phrase-based model, \(X\) is a non-terminal symbol, \(\gamma\) is a sequence of non-terminals and source terminals and \(\alpha\) is a sequence of non-terminals and target terminals. Symbol \(\sim\) is a one-to-one correspondence for the non-terminals appeared in \(\gamma\) and \(\alpha\). To build an interpolated Kneser-Ney language model [18] on the target side of the training data, SRILM [19] toolkit is used. Parameters are initialized as follows: language model parameter is initialized to 1, word penalty is initialized to -2.8 and the other parameters are initialized to 0. All the parameters have default values in Joshua decoder. Finally Joshua decoder [20] decodes the best translation with the log-linear method. Joshua decoder is used to decode the test set. Joshua decoder is an implementation of the CKY+ algorithm [21] and implements scope-3 filtering [22] and uses cube pruning [23] to reduce parsing complexity [20] when filtering grammars to test sets. The decoder is employed to produce the k-best translations for each sentence of the test set. Decoding algorithm maintains cubic time parsing complexity (in the sentence length).

4. Results and discussion

This section presents the experiments and the
The model needs parallel corpus which consists of unedited corpus and the edited one. A dataset with these aligned corpora is not available for Persian language. Two criteria are specified for creating a dataset for this special purpose: First criterion states that space and half-space characters must be denoted as two different symbol characters in the corpora. The second criterion is to create a dataset of parallel corpora in which unedited multi-part words are placed in one side and edited multi-part words are placed in the other side. In the edited side of parallel corpora, spaces between the parts of the multi-part words are replaced by half-spaces. Therefore, a dataset is created based on the two criteria and it is publicly available for other researchers. The model needs dataset especially for evaluation step. The evaluation set must consist of two sets: one for tuning parameters of the model, and the other one for validation experiments. A tuning set is created and used to set the parameters of model in order to use minimum error rate training in the training step.

As shown in table 2, the words such as “وقت یکه” are edited successfully, because the training set includes these words. As it is shown in table 2, the words like “می خدمت” are edited successfully, however, it is not exactly included in the training set. This is the ability of the proposed method to edit the words which are not exactly included in the training dataset. The training dataset contains the words with similar structure with sufficient frequency. Therefore, the proposed method can model the co-occurrence of parts of the multi-part...
words. In more details, one can see that the word “یحوصله” is not edited. As shown in table 3, by increasing the frequency of similar words like “یحوصله” and “یحوصله” makes it possible for the proposed method to edit “یحوصله” correctly.

Table 2. A sample of evaluation output.

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>وست و نکان دادن سر نمادی از فردی بی حوصله است</td>
<td>وست و نکان دادن سر نمادی از فردی بی حوصله است</td>
</tr>
<tr>
<td>لیب‌ها رو به سمت بالا کشیده می‌شود که این شانش می‌دهد، ترد</td>
<td>لیب‌ها رو به سمت بالا کشیده می‌شود که این شانش می‌دهد، ترد</td>
</tr>
<tr>
<td>واقعا حوصله است. زمانی که از نشانی فردی حوصله می‌کنم</td>
<td>واقعا حوصله است. زمانی که از نشانی فردی حوصله می‌کنم</td>
</tr>
<tr>
<td>به یکی که در ناکامی‌های فردی حوصله می‌کنم</td>
<td>به یکی که در ناکامی‌های فردی حوصله می‌کنم</td>
</tr>
</tbody>
</table>

Table 3. Evaluation in the case of increasing the frequency of words with the same structure in the training dataset.

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>وست و نکان دادن سر نمادی از فردی بی حوصله است</td>
<td>وست و نکان دادن سر نمادی از فردی بی حوصله است</td>
</tr>
<tr>
<td>لیب‌ها رو به سمت بالا کشیده می‌شود که این شانش می‌دهد، ترد</td>
<td>لیب‌ها رو به سمت بالا کشیده می‌شود که این شانش می‌دهد، ترد</td>
</tr>
<tr>
<td>واقعا حوصله است. زمانی که از نشانی فردی حوصله می‌کنم</td>
<td>واقعا حوصله است. زمانی که از نشانی فردی حوصله می‌کنم</td>
</tr>
<tr>
<td>به یکی که در ناکامی‌های فردی حوصله می‌کنم</td>
<td>به یکی که در ناکامی‌های فردی حوصله می‌کنم</td>
</tr>
</tbody>
</table>

Therefore, if the sufficient number of the multi-part words with the similar structure exist in the training set, the multi-part word would be edited even the word is unseen in the training set.

There are some multi-part words, where each part can be considered as an independent word such as “یهم” and “یهم”. If maximum entropy POS tagger [24] is used to train the tags, it cannot perform efficiently. Since maximum entropy approach edits the spacing by using maximum co-occurrence of space and half-space between the parts and since the maximum co-occurrence does not have linguistic information to edit spacing, the approach is not efficient to edit spacing. If the co-occurrence of half-space after “یهم” is more than the co-occurrence of space, the space is edited to half-space while the word “یهم” can be considered as an independent word.

Therefore, correct spacing would not be achieved by just relying on the co-occurrence of space and half-space characters between the parts of multi-part words, while in the proposed approach, spacing in multi-part words can be edited successfully because of using linguistic information.

The approach is evaluated using False Positive (FP), False Negative (FN), Precision (P) and Recall (R) measures. Recall (R) and Precision (P) are calculated using the following equations.

Recall(R)= (Number of correct edited multi-part words)/(Total number of multi-part words in the text) (6)

Precision (P)= (Number of correct edited multi-part words)/(Number of edited words) (7)

Recall is also considered to be the accuracy score of the approach by calculating number of correct edited multipart words against the total number of multi-part words in the corpus. Precision is also considered to be the accuracy score of the approach by calculating number of correct edited multi-part words against the total number of edited words which are edited by the approach. The accuracy rate is computed with the average of four different created test sets. In the proposed approach, recall and precision are obtained 92% and 98%, respectively. The score of false positive and false negative are 1.8% and 3%, respectively.

Another measure used to evaluate the efficiency of the proposed method is BLEU [25]. BLEU is not an error rate but an accuracy measure [26] and it discovers the best scoring result as follows.

\[ P(w_1, w_2, ..., w_T) = P(w_1) P(w_2|w_1) P(w_3|w_1, w_2) ... P(w_T|w_1, ..., w_{T-1}) \]

where, \( w_1, ..., w_T \) is a sentence and \( w_i \) is the i-th word of sentence.

BLEU score of the proposed method reaches 0.91.

5. Conclusion

In this paper, a statistical approach is introduced to edit Persian text focusing on spacing in Persian multi-part words. The paper employs statistical machine translation which translates one language
into another. The proposed approach utilizes this ability to edit Persian text. Thus, the proposed approach employs parallel corpora in which unedited multi-part words are considered as source language and space-edited multi-part words are considered as destination language. Since no standard dataset exists in literature, three Persian parallel corpora is prepared to meet the needs; one for train, one for tune and one for test. To align the created parallel corpora, the proposed method employs a fertility-based IBM model and calculates the parameters of probabilistic distributions and extracts linguistic information with Synchronous Context-Free Grammars (SCFG) of hierarchical phrase-based model. In evaluation phase, a syntax-based decoder is used to decode different created test sets in this paper. Based on this model, multi-part words are edited efficiently even the words are not exactly trained in the training set provided that the same word structure is trained in the training set. Furthermore, the experimental validation shows that the proposed method can edit spacing in multi-part words with a desired result.

References
Conference on Spoken Language Processing, Denver, Colorado, USA, 2002.


یک مدل جدید برای ویرایش کلمات چندبخشی فارسی براساس ترجمه ماشینی آماری
مرتضی زاهدی و آرزو ارجمندزاده
دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی شاهرود، ایران.
ارسال 70/70/4702؛ پذیرش 72/70/4702
چکیده
اجزای کلمات چندبخشی در زبان انگلیسی با استفاده از خط ربط از هم جدا می‌شوند. در زبان فارسی برای جداکردن اجزاء، کلمات چندبخشی و در عین حال حفظ یکپارچگی اجزاء به‌عنوان یک کلمه واحد، از نیمفاصل استفاده می‌شود. در بسیاری از موارد به نادرستی بین اجزاء یک کلمه چندبخشی قابل قرار گیرد که این باعث بوجود آمدن مشکلات در پردازش متن فارسی و همچنین باعث کاهش خوانایی متن می‌شود. در این مقاله تأکید گردیده که در زبان فارسی برای اجرای فاصله شده‌کردن، باید استفاده از یک پارادایم ترجمه ماشینی آماری که بهترین پاسخگویی به مشکلات این زبان داشته باشد، بررسی شود.
در این بستر، پارادایم ترجمه ماشینی ابزاری است که می‌تواند کلمات چندبخشی را به‌طور صحیح ویرایش کند.
کلمات کلیدی: کلمات چندبخشی فارسی، قوانین فاصله‌گذاری، ترجمه ماشینی آماری، پیکره مواری، عبارات سلسله مراتبی، مدل IBM مبتنی بر باروری، ترجمه مبتنی بر نحو.