Journal of AI and Data Mining
Vol 4, No 1, 2016, 27-34 ***10.5829/idosi.JAIDM.2016.04.01.04***

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-joiner 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 doses not obey standard word form and each part will be incorrectly considered as a separate word such as, "بی شمار", "هیچ گاه" and "حاصل 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

with the part of speech tags. Mahmoudi et al. [3] 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-joiner 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:

غیرقابل، بی‌حوصله، پائین‌تر، جریمه‌های، رقابت‌های، ازدست نمی‌دهد، می‌شود، تصور نموده‌اند، می‌بایست

Half-space is a character with zero-width non-joiner 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 “زبان‌شناسی” and “می‌شود″ 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

P(e)P(f |e) as large as possible, equation (3) is presented [6,8,9].

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, \mathrm{P}(e|f) \quad (1)$$

$$\mathrm{P}(e \mid f) = (\mathrm{P}(e)\ \mathrm{P}(f|e))/(\mathrm{P}(f)) \quad (2)$$

$$\hat{e} = \operatorname{argmax}_e\, \mathrm{P}(e)\, \mathrm{P}(f|e) \quad (3)$$

P(e) is the prior probability and P(f |e) is the conditional probability of target language word with given the source language word and $\hat{e}$ is the maximum probability product of P(f)P(e| f).

SMT requires a parallel corpus to extract linguistic information for each language pair. In first step, SMT assigns translation probability for each parallel word with aid of the IBM model [6] which is used as the word alignment method in this paper. Brown et al. [6] proposes five statistical models for the translation process and the computational complexity increases through going from Model 1 to Model 5 while it is closer to human language and requires additional parameters [10].

3. Materials and methods

3.1. Fertility-based IBM model and hierarchical phrase-based model

IBM Model 3 [6] consists of three parameters: lexicon model parameter, fertility model parameter and distortion model parameter. The generative story of the IBM model 3 focuses on training which is based on the concept of fertilities:

Given a vector alignment of a source sentence $\mathbf{a}_1^J$, the fertility of target word i expresses the number of source words aligned to it [11].

$$\Phi_i\left(\mathbf{a}_1^J\right) = \sum_{j:a_j=i} 1 \quad (4)$$

It omits the dependency on $\mathbf{a}_1^J$ (and defining P(j | 0)=1), the probability is expressed as follows.

$$\mathrm{P}\left(f_1^J, a_1^J \middle| e_1^I\right) = \mathrm{P}(\Phi_0 | J) \cdot \prod_{i=1}^{I}[\Phi_i!\, \mathrm{P}(\Phi_i | e_i)] \cdot \prod_j \left[\mathrm{P}\left(f_i \middle| e_{a_j}\right) \cdot \mathrm{P}(j \mid a_j)\right] \quad (5)$$

For each foreign input word f, it factors on the fertility probability P(Φ_i |f_i) . The factorial Φ_i! stems from the multiple tableaux for one alignment, if Φ_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 "زاده". So, the space character between the two parts must be edited into half-space character.

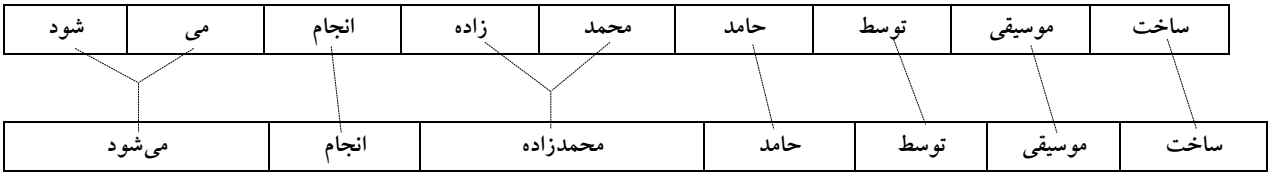

Figure1. Word alignment in Persian language.

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 "*" 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
به*گوشش*خورد:*«هرکس*از*ما*کمکی*بخواهد*ما*به*او*کمک*می*کنیم،* ولی*اگر*کسی*بی*نیازی*بورزد*و*دست*حاجت*پیش*مخلوقی*دراز*نکند،* خداوند*او*را*بی*نیاز*می*کند.*آن*روز*چیزی*نگفت.*و*به*خانه*خویش* برگشت.*	به*گوشش*خورد:*«هرکس*از*ما*کمکی*بخواهد*ما*به*او*کمک*می&کنیم،* ولی*اگر*کسی*بی&نیازی*بورزد*و*دست*حاجت*پیش*مخلوقی*دراز*نکند،* خداوند*او*را*بی&نیاز*می&کند.*آن*روز*چیزی*نگفت.*و*به*خانه*خویش* برگشت.*
باز*با*هیولای*مهیب*فقر*که*هم*چنان*بر*خانه*اش*سایه*افکنده*بود*روبرو* شد،*ناچار*روز*دیگر*به*همان*نیت*به*مجلس*رسول*اکرم*حاضر*شد،*آن* روز*هم*همان*جمله*را*از*رسول*اکرم*شنید:*«هرکس*از*ما*کمکی*بخواهد *ما*به*او*کمک*می*کنیم،*ولی*اگر*کسی*بی*نیازی*بورزد،*خداوند*او*را* بی*نیاز*می*کند.این*دفعه*نیز*بدون*این*که*حاجت*خود*را*بگوید،*به*خانه* خویش*برگشت،*چون*خود*را*هم*چنان*در*چنگال*فقر*ضعیف*و*بیچاره*و* ناتوان*می*دید،*برای*سومین*بار*به*همان*نیت*به*مجلس*رسول*اکرم*رفت،* باز*هم*لب*های*رسول*اکرم*به*حرکت*آمد*و*با*همان*آهنگ*که*به*دل* قوت*و*به*روح*اطمینان*می*بخشید*همان*جمله*را*تکرار*کرد*.	باز*با*هیولای*مهیب*فقر*که*هم&چنان*بر*خانه&اش*سایه*افکنده*بود*روبرو* شد،*ناچار*روز*دیگر*به*همان*نیت*به*مجلس*رسول&اکرم*حاضر*شد،*آن* روز*هم*همان*جمله*را*از*رسول&اکرم*شنید:*«هرکس*از*ما*کمکی*بخواهد *ما*به*او*کمک*می&کنیم،*ولی*اگر*کسی*بی&نیازی*بورزد،*خداوند*او*را* بی&نیاز*می&کند.این*دفعه*نیز*بدون*این&که*حاجت*خود*را*بگوید،*به*خانه* خویش*برگشت،*چون*خود*را*هم&چنان*در*چنگال*فقر*ضعیف*و*بیچاره*و* ناتوان*می&دید،*برای*سومین*بار*به*همان*نیت*به*مجلس*رسول&اکرم*رفت،* باز*هم*لب&های*رسول*اکرم*به*حرکت*آمد*و*با*همان*آهنگ*که*به*دل* قوت*و*به*روح*اطمینان*می&بخشید*همان*جمله*را*تکرار*کرد*.

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 $P_w(\gamma|\alpha)$ parameter and $P_w(\alpha \mid \gamma)$ 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 $X \rightarrow <\gamma, \alpha, \sim, w>$ in hierarchical phrase-based model, X is a non-terminal symbol, γ is a sequence of non-terminals and source terminals and α is a sequence of non-terminals and target terminals. Symbol ~ is a one-to-one correspondence for the non-terminals appeared in γ and α. 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

results of created test sets. 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.

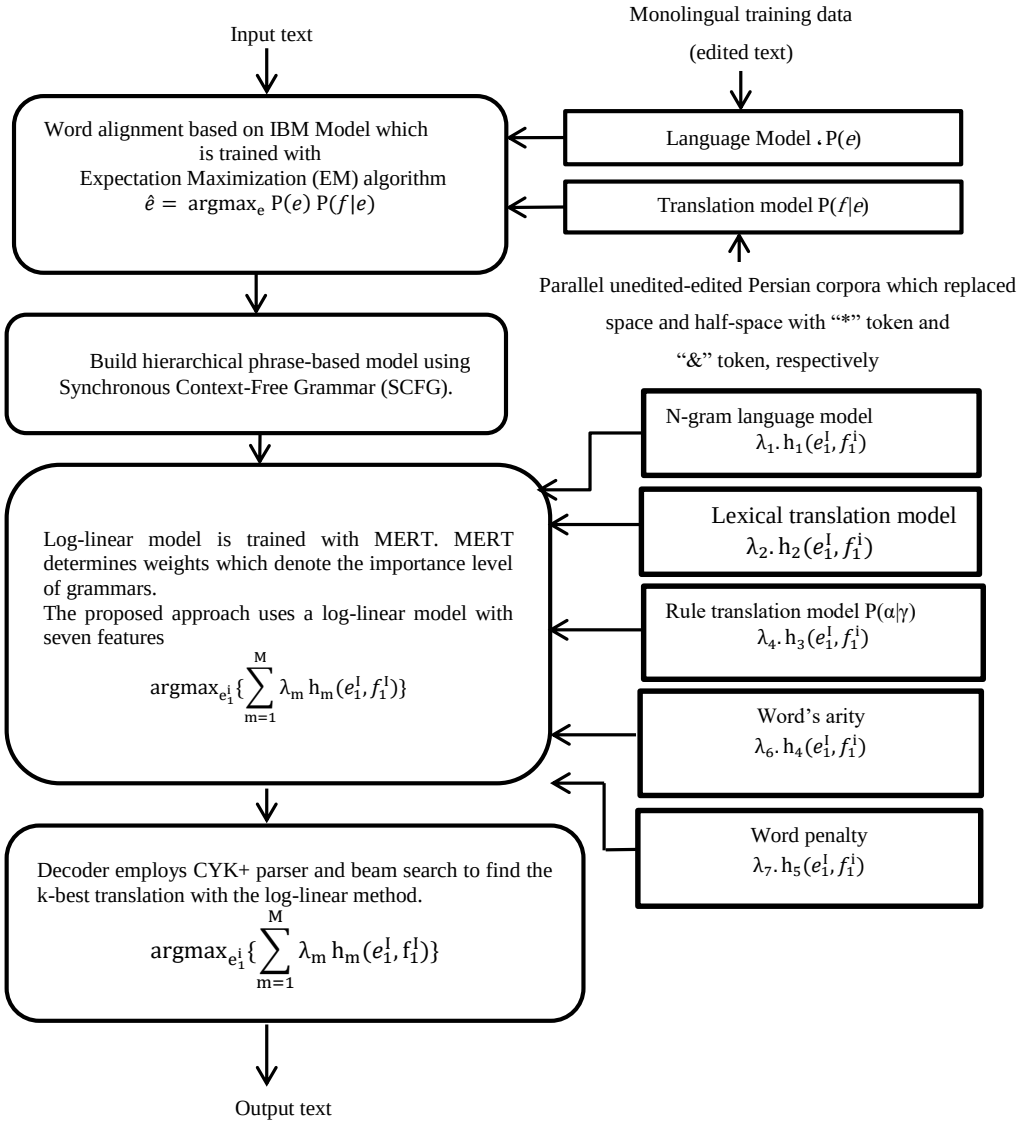

Figure 2. Graphical representation of the proposed method.

As shown in table 2, the words such as "وقتی‌که", "گونه‌ای", "می‌شود", "لب‌ها", "زمانی‌که", "می‌دهد" are edited successfully, because the training set includes these words. As it is shown in table 2, the words like "می‌خندیم" is 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 "بی‌حوصله" is not edited. As shown in table 3, by increasing the frequency of similar words like "بی‌دغدغه", "بی‌امان" and "بی‌همتا" makes it possible for the proposed method to edit "بی‌حوصله" correctly.

Table 2. A sample of evaluation output.

Input	Output
وقتی که واقعا می خندیم عضله گونه ای بزرگ و منقبض می شود و لب ها رو به سمت بالا کشیده می شود که این نشان می دهد، فرد واقعا خوشحال است. زمانی که از دیدن فردی خوشحالیم، سرمان را بالا می گیریم و برعکس سر پایین نشان از ناراحتی از حضور اوست وتکان دادن سر نمادی از فرد بی حوصله است.	وقتی‌که واقعا می‌خندیم عضله گونه‌ای بزرگ و منقبض می‌شود و لب‌ها رو به سمت بالا کشیده می‌شود که این نشان می دهد، فرد واقعا خوشحال است. زمانی‌که از دیدن فردی خوشحالیم، سرمان را بالا می‌گیریم و برعکس سر پایین نشان از ناراحتی از حضور اوست و تکان‌دادن سر نمادی از فرد بی حوصله است .

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

Input	Output
وقتی که واقعا می خندیم عضله گونه ای بزرگ و منقبض می شود و لب ها رو به سمت بالا کشیده می شود که این نشان می دهد، فرد واقعا خوشحال است. زمانی که از دیدن فردی خوشحالیم، سرمان را بالا می گیریم و برعکس سر پایین نشان از ناراحتی از حضور اوست وتکان دادن سر نمادی از فرد بی حوصله است.	وقتی‌که واقعا می‌خندیم عضله گونه‌ای بزرگ و منقبض می‌شود و لب‌ها رو به سمت بالا کشیده می‌شود که این نشان می‌دهد، فرد واقعا خوشحال است. زمانی که از دیدن فردی خوشحالیم، سرمان را بالا می‌گیریم و برعکس سر پایین نشان از ناراحتی از حضور اوست و تکان دادن سر نمادی از فرد بی‌حوصله است .

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 "به" and "ویژه" in "به‌ویژه". 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 "به" is more than the co-occurrence of space, the space is edited to half-space while the word "به" 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.

$$P(w_1, w_2, \dots, w_T) = P(w_1)\, P(w_2|w_1) P(w_3|w_1 w_2) \dots P(w_T|w_1, \dots, w_{T-1}) \quad (8)$$

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

[1] Kashefi, O., Sharifi, M. & Minaei-Bidgoli, B. (2012). A Novel String Distance Metric for Ranking Persian Respelling Suggestions. Natural Language Engineering (NLE), Cambridge University Press, United Kingdom, vol. 2, no. 19, pp.259–284.

[2] Shamsfard, M., Sadat Jafari, H. & Ilbeygi, M. (2010). STeP-1: A Set of Fundamental Tools for Persian Text Processing. The 7th International Conference on Language Resources and Evaluation, Valletta, Malta, 2010.

[3] Mahmoudi, A., Faili, H. & Arabsorkhi, M. (2013). Modeling Persian Verb Morphology to Improve English-Persian Machine Translation. The 12th Mexican International Conference on Artificial Intelligence, Mexico City, Mexico, 2013.

[4] Rasooli, M. S., Kholy, A. E. & Habash, N. (2013). Orthographic and Morphological Processing for Persian to English Statistical Machine Translation. The 6th International Joint Conference on Natural Language Processing, Nagoya, Japan, 2013.

[5] Bijankhan, M., Sheykhzadegan, J., Bhrani, M. & Ghayoomi, M. (2010). Lessons from Building a Persian Written Corpus: Peykare. Language Resources and Evaluation, vol. 54, no. 2, pp. 143-164.

[6] Brown, P .E., Della Pietra, S. A., Della Pietra, V. J. & Mercer, R.L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics - Special Issue on Using Large Corpora: II, vol. 19, no. 2, pp. 263-311.

[7] Chiang, D. (2005). A Hierarchical Phrase-based Model for Statistical Machine Translation. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan, 2005.

[8] Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Mercer, R. L. & Roossin, P. S. (1990). A Statistical Approach to Machine Translation. Computational Linguistics, vol. 16, no. 2, pp. 79-85.

[9] Berger, A. L., Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Gillett, J. R., Lafferty, J. D., Mercer, R. L., Printz, H. & Ures, L. (1994). The Candide System for Machine Translation. The workshop on Human Language Technology, Stroudsburg, USA, 1994.

[10] Kohen, Ph. (2010). Statistical Machine Translation. Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, Sao Paulo, Delhi, Dubai, Tokyo: Cambridge University Press.

[11] Schoenemann, T. (2010). Computing Optimal Alignments for the IBM-3 Translation Model. The 14th Conference on Computational Natural Language Learning, Uppsala, Sweden, 2010.

[12] Och, F. J. & Ney, H. (2000). Improved Statistical Alignment Models. The 38th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, 2000.

[13] Och, F. J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, vol. 29, no. 1, pp. 19-51.

[14] Weese, J., Ganitkevitch, J., Callison-Burch, Ch., Post, M. & Lopez, A. (2011). Joshua 3.0: Syntax-based Machine Translation with the Thrax Grammar Extractor. The 6th Workshop on Statistical Machine Translation, Stroudsburg, PA, USA, 2011.

[15] Zollmann, A. & Venugopal, A. (2006). Syntax Augmented Machine Translation via Chart Parsing. The Workshop on Statistical Machine Translation, Stroudsburg, USA, 2006.

[16] Zaidan, O. F. (2009). Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. The Prague Bulletin of Mathematical Linguistics, Czech Republic, 2009.

[17] Och, F. J. (2003). Minimum Error Rate Training for Statistical Machine Translation. The 41st Annual Meeting on Association for Computational Linguistics, Stroudsburg, USA, 2003.

[18] Heafield, K., Pouzyrevsky, I., Clark, J. H. & Kohen, Ph. (2013). Scalable Modified Kneser-Ney Language Model Estimation. The 51st Annual Meeting on Association for Computational Linguistics, Sofia, Bulgaria, 2013.

[19] Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. The 7th International

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

[20] Post, M., Ganitkevitch, J., Orland, L., Weese, J. & Cao, Y. (2013). Joshua 5.0: Sparser, Better, Faster, Server. The Eighth Workshop on Statistical Machine Translation, Sofia, Bulgaria, 2013.

[21] Chappelier, J. & Rajman, M. (1998). A Generalized CYK Algorithm for Parsing Stochastic CFG. The 1st Workshop on Tabulation in Parsing and Deduction, Paris, France, 1998.

[22] Hopkins, M. & Langmead, G. (2010). SCFG Decoding without Binarization. The 2010 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Cambridge, USA, 2010.

[23] Chiang, D. (2007). Hierarchical Phrase-based Translation. Computational Linguistics, vol. 33, no. 2, pp. 201–228.

[24] Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. The 1996 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Philadelphia, USA, 1996.

[25] Papineni, K., Roukos, S., Ward, T. & Wei-Jing Zhu (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. The 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, USA, 2002.

[26] Gonzalez, J. (2012). A Finite State Approach to Phrase-based Statistical Machine Translation. The 10th International Workshop on Finite State Methods and Natural Language Processing, Spain, 2012.

نشریه هوش مصنوعی و داده‌کاوی

دانشگاه شاهرود

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، شاهرود، ایران.

ارسال ۲۰۱۴/۰۷/۰۷؛ پذیرش ۲۰۱۵/۰۷/۰۴

چکیده:

اجزاء 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 نیم‌فاصله استفاده می‌شود. در بسیاری از موارد به نادرستی بین اجزاء یک کلمه چندبخشی فاصله قرار می‌گیرد که این باعث بوجود آمدن مشکلاتی در پردازش متن فارسی و همچنین باعث کاهش خوانایی متن می‌شود. در این مقاله روشی ارائه شده‌است که با استفاده از آن می‌توان فاصله میان اجزاء کلمات چندبخشی را با نیم‌فاصله ویرایش کرد. در روش ارائه‌شده، پارادایم ترجمه ماشینی آماری برای ویرایش فاصله میان اجزاء کلمه چندبخشی به‌کارگرفته شده‌است. در ترجمه ماشینی آماری از یک پیکره موازی برای استخراج پارامترهای آماری و اطلاعات زبانی استفاده می‌شود. به‌طوری‌که در سمت مبدا پیکره موازی، متن زبان مبدا و در سمت هدف آن متن زبان مقصد قرار دارد. پیکره موازی که در این مقاله ایجاد شده به این‌صورت است که در سمت مبدا متنی با کلمات چندبخشی که بین اجزاء آن فاصله قرار دارد، آمده‌است و در سمت هدف این فاصله‌ها به نیم‌فاصله ویرایش شده‌اند. نتایج نشان‌دهنده این است که روش ارائه‌شده می‌تواند با میزان دقت چشمگیری فاصله میان کلمات چندبخشی را به نیم‌فاصله ویرایش کند.

کلمات کلیدی: کلمات چندبخشی فارسی، قوانین فاصله‌گذاری، ترجمه ماشینی آماری، پیکره موازی، عبارات سلسله مراتبی، مدل IBM مبتنی‌بر باروری، ترجمه مبتنی‌بر نحو.