

Joint Persian Word Segmentation Correction and Zero-Width Non-Joiner Recognition Using BERT

Ehsan Doostmohammadi*, Minoo Nassajian*, Adel Rahimi†

*Sharif University of Technology, Tehran, Iran

†Dathena Science Pte. Ltd., Singapore

{e.doostm72,m.nassajian2016}@student.sharif.edu
adel.rahimi@dathena.io

Abstract

Words are properly segmented in the Persian writing system; in practice, however, these writing rules are often neglected, resulting in single words being written disjointedly and multiple words written without any white spaces between them. This paper addresses the problems of word segmentation and zero-width non-joiner (ZWNJ) recognition in Persian, which we approach jointly as a sequence labeling problem. We achieved a macro-averaged F₁-score of 92.40% on a carefully collected corpus of 500 sentences with a high level of difficulty.

1 Introduction

People who have worked with real-world data in Persian natural language processing (NLP) are familiar with the Persian writing system’s problematic properties, the most important of which are word segmentation and ZWNJ, both consequences of a phenomenon called “orthographic ligature”. In Persian, white space character (U+0020) is used to segment words in a sentence, and ZWNJ (U+200C) is used between free morphemes in compound words (or combining forms), and also between one or several free morphemes and one or several bound morphemes to stop joining letters from connecting into a ligature (Moghaddam, 2007). Although there is a set of rules regarding the use of white space and ZWNJ set by the regulatory body of the Persian language (Academy of Persian Language and Literature, 2015), only a few people follow them in writing formal Persian, let alone the informal language.

Some common errors regarding the use of white space and ZWNJ are as follows: (1) not segmenting words where the last letter of the first word is a non-joiner character (e.g., *va_āmar[-e]_marg|omir_behtar_az_xeyli|hāst*¹ “and the mortality rate is better than many”); (2) using ZWNJ instead of white space which could be categorized as (a) excessive use of ZWNJ (e.g., for separating first and last names) and (b) ZWNJ key hit mistakenly as in many cell phone keyboards the space key and the ZWNJ key are adjacent to each other; (3) using space instead of ZWNJ (e.g. *mi konam* “I do”); (4) using neither of them (e.g. *mikonam* “I do”); and many other spontaneous errors.

In comparison to European languages, such as English and German, word segmentation in Asian languages like Japanese, Chinese (Li et al., 2019), and Thai (Aroonmanakun, 2002; Tesprasit et al., 2003), is more complicated, because space is not specifically used as an orthographic word boundary delimiter (Khan et al., 2018b). For instance, in Vietnamese, space is not only used to separate words, but it is also applied to separate syllables (that can be considered as a meaningful word or as a part of multi-syllable words) that make up words (Huyen et al., 2008). The problem in Persian is not similar to that of the abovementioned languages, but quite identical to Urdu (Durrani and Hussain, 2010; Lehal, 2010; Rashid and Latif, 2012; Bin Zia et al., 2018; Khan et al., 2018a), as they both are Indo-Iranian languages and their writing systems are derived from the Arabic script. In Persian, words are properly segmented in theory, but these rules are not always followed in practice. These kinds of writing system characteristics result in difficulties in text processing, e.g., sequence labeling. According to the previous research, word segmentation can improve the results of other NLP tasks like information retrieval (Foo

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

¹In this example there was no white space or ZWNJ. We added `_` where white space and `|` where ZWNJ should have been.

and Li, 2004), machine translation (Xu et al., 2004), information extraction (Peng and Dredze, 2016), dependency parsing (Nguyen, 2018), etc.

In this paper, we address the problems of word segmentation and ZWNJ correction in Persian using sequence labeling models and achieve results that are quite promising and could pave the way for an effective solution for real-world situations. To be sure of this, we gathered a corpus of 500 sentences with a high degree of difficulty regarding the correctness of word segmentation and ZWNJs, and evaluated the models' performance on it. After reviewing the related work, we discuss the data used for training, validating, and testing the models in §3, the methodology and the experimental settings in §4, and the results in §5. We then conclude the paper and suggest some ideas for future work.

2 Related Work

Segmentation techniques could be classified into different categories, such as rule-based methods (Riaz, 2010; Khan et al., 2018a), hybrid methods (Nakagawa and Uchimoto (2007) and Pham et al. (2009) use a hidden Markov model with some hand-crafted rules), and machine learning and deep learning methods (Sassano (2002) and Nguyen et al. (2006) use conditional random fields and support vector machines, and Ma et al. (2018) use a bidirectional recurrent neural network model). In addition to supervised learning (like Zhang et al. (2016), Lyu et al. (2016), etc.), there are studies on semi-supervised (Fujii et al., 2017; Wang et al., 2019; Yang and Vozila, 2014) and unsupervised word segmentation (Mochihashi et al., 2009; Seeha et al., 2020; Zhang et al., 2018; Bui et al., 2019). As for the latest studies, we can mention Ke et al. (2020), Nguyen (2018), and Yan et al. (2020) that use BERT and multitask learning methods to segment words in Chinese and Vietnamese languages.

In Persian, there are few studies conducted on preprocessing toolkits, some of which also include modules for space and ZWNJ correction. As an early work, we can mention Shamsfard et al. (2009) that designed a Persian text preprocessing tool called STeP-1 to tokenize texts, check spellings, and analyze words morphologically. To design the tokenizer, they use dictionary-based and rule-based methods to recognize the correct places of space and ZWNJ, and report a performance of 86.6%. As another work, Sarabi et al. (2013) design a toolkit for Persian text processing in four lexical, morphological, syntactic, and semantic levels. Space and ZWNJ positions are checked in the first level using dictionary-based and rule-based methods, achieving a precision of 95%. Additionally, Hazm², an open-source preprocessing library, performs ZWNJ correction using regular expressions and a list of valid stems. Parsivar (Mohtaj et al., 2018), the most powerful Persian preprocessing tool, applies rule-based and statistical methods to determine the correct places of ZWNJ and space, respectively. They trained a naive Bayes model on the 10 million word Bijankhan corpus (Bijankhan et al., 2011) with the IOB tagging scheme to find word boundaries, achieving an F_1 -score of 96.5%. As for the latest work, Panahandeh and Ghanbari (2019) uses an N-gram language model and a rule-based method to correct space and ZWNJ between compound words, respectively. They report an F_1 -score of 81.94% for space correction.

3 Data

The 10 million word Bijankhan corpus (Bijankhan et al., 2011) was used in this research, which is a cleanly tokenized corpus with fine-grained part-of-speech tags (which were not used here). This corpus contains 10,437,194 words or 38,971,131 characters. We chose this corpus since (a) their word segmentation is clean and their approach is practical, (b) the size of the corpus is enormous (~ 39 M characters), and (c) the corpus comprises many different topics, including news articles, literary prose and poems, informal dialogues, etc. All white spaces in a single token were converted to ZWNJs and the characters in the corpus were normalized as well. Then, the corpus was split into three parts: the first 10% for testing, the second 10% for validation, and the rest for training. We also collected a test corpus of 500 sentences from twitter, news broadcasting websites, and discussion forums to have real-world data with real white space and ZWNJ errors. This data was deliberately meant to be a difficult test corpus with numerous extreme cases (and also easy and completely correct ones). This corpus contains 16,574 words or 93,355 characters.

²<https://github.com/sobhe/hazm>

4 Methodology

We see the word segmentation correction and the ZWNJ recognition tasks as one problem and train a single model to perform these two tasks jointly. We approached the task as a sequence labeling problem, i.e., mapping each character to a tag space of size 3. The tag is 0 when there is no white space or ZWNJ, 1 when there is white space, and 2 when there is ZWNJ after the corresponding character. Two types of models were trained for this purpose:

1. **CRF**: a conditional random field (CRF) model (Lafferty et al., 2001) implemented using `sklearn-crfsuite` (Korobov, 2015; Okazaki, 2007), with the input features of the focus, 5 previous and 5 following characters, and four character-based Boolean features indicating whether the focus character is `is_first` and `is_last` character of the sentence, and also if the character is `is_joiner`, and whether it is `is_digit`. All the white space and ZWNJ characters were stripped from the input texts. The L1 and L2 regularization coefficients were set to 0.1 and the max iteration argument to 100.
2. **BERT**: The main bidirectional encoder representations from Transformers (BERT) model (Devlin et al., 2018), plus a fully-connected network mapping to the tag space. The learning rate was set to $2e - 5$ and the batch size to 10. Adam (Kingma and Ba, 2014) was used for optimizing the weights with cross-entropy as the loss function. As for the pre-trained weights, the `multilingual_cased` model was used. We have followed the recommended settings for sequence labeling, which is to calculate loss only on the first part of each tokenized word. The implementation was done using PyTorch (Paszke et al., 2019) and HuggingFace’s Transformers (Wolf et al., 2019) libraries. The input texts were fed into the model in two different settings:
 - (a) All the white space and ZWNJ characters were stripped from the input texts, similar to the CRF model above. The model has to figure out the position of white space and ZWNJ characters in sentences from scratch. We refer to this model as `BERTa`.
 - (b) The white space and ZWNJ characters were kept, but some noise was introduced to the input data in the following manner for each sentence in Bijankhan corpus:
 - i. $r_1 \times l$ of the ZWNJs were changed to white space characters, where l is the length of the sentence (in characters) and r_1 is a uniform random variable on $(0., .15)$.
 - ii. $r_2 \times l$ of the white spaces after non-joiner characters were removed, where l is the length of the sentence (in characters) and r_2 is a uniform random variable on $(0., .2)$.
 - iii. $r_3 \times l$ of the remaining characters were changed in the following manner, where l is the length of the sentence (in characters) and r_3 is a uniform random variable on $(0., .05)$:
 - A. `replace(c, rand(null, ZWNJ))`, where the randomly chosen character c was white space;
 - B. `replace(c, rand(null, space))`, where the randomly chosen character c was ZWNJ;
 - C. `replace(c, concat(c, rand(ZWNJ, space)))`, where the randomly chosen character c was followed neither by a white space nor a ZWNJ, otherwise, remove the following white space or ZWNJ.

The noise ranges were chosen based on our observation of real-world errors; hence this scenario is closer to real-world situations. The corresponding output tags of the model for the input white space and ZWNJ character were masked and ignored in calculating the loss and performance measures. We refer to this model as `BERTb`.

We also experimented with Parsivar tool, introduced in §2, to rectify white spaces and ZWNJs using its `Normalizer` (with the `statistical_space_correction` argument set to `True`) and `SpellCheck` modules. As the latter module might add or remove characters other than white space and ZWNJ, the differences between the two strings were found using Python’s `difflib` library, and placeholder characters were added where needed to make the strings the same length and make their comparison possible. This work’s performance measures are precision, recall, and F_1 -score for each class and macro-averaged F_1 -score of all of them.

5 Results and Analysis

The results of the abovementioned methods on the test set of Bijankhan corpus and the 500 sentence corpus are shown in Table 1. The BERT models outperform the other methods by a large margin, BERT_b standing on the top by 1.33% F₁-score more than BERT_a, 28.65% more than the CRF, 26.73% more than Parsivar, and 9.55% more than the baseline on the 500 sentence corpus. The baseline simply indicates the correctness of word segmentation and ZWNJs in the 500 test corpus (i.e., naturally occurring errors in the real-world data) when compared to its corrected pair. The results also show that Parsivar and the CRF model not only do not increase word segmentation and ZWNJs correctness, but also add more errors to the data.

	Bijankhan Test Set				500 Sentence Corpus			
	Class F ₁			Avg. F ₁	Class F ₁			Avg. F ₁
	0	1	2		0	1	2	
Baseline					0.9823	0.9336	0.5697	0.8285
Parsivar					0.9561	0.7785	0.2355	0.6567
CRF	0.8923	0.6369	0.5664	0.6985	0.8826	0.6080	0.4217	0.6375
BERT_a	0.9937	0.9779	0.9286	0.9667	0.9832	0.9488	0.8000	0.9107
BERT_b	0.9963	0.9886	0.9593	0.9814	0.9856	0.9592	0.8272	0.9240

Table 1: The F₁-score for each class and the macro-averaged F₁-scores of all the classes for each of the methods on Bijankhan test set and the 500 sentence corpus. The best result in each column is in bold.

The errors in BERT_a and BERT_b are more or less similar and can be categorized into the following groups: (a) ZWNJ before an intrusive *y* between a vowel and an *ezafe* (e.g., *xāney[e]* “[the] house of”), which is simply because this writing style is not used in Bijankhan corpus (and is also fixable as words with *ezafe* are labeled as GEN in Bijankhan corpus); (b) underscores and hash characters in hashtags, which is again because the model has not seen it in the training data; (c) out of vocabulary words, such as *tu’iter* “twitter” and *dāmāš* “damash (a soccer team name)”; (d) digits occasionally sticking to the word before or after them; (e) informal words, such as the clitic *e* “is”, which is the short form of *ast* “is”; (f) words or phrases with different syntactic roles, such as *xaridāri|šode* “bought (adjective)” and *xaridāri šode* “bought (verb)”; (g) typos and uncommon spellings. The confusion matrix also reveals that ZWNJ is mostly mistaken with the 0 class, and not with the 1 class, which results in joiner characters connecting into a ligature and make the text difficult to read. There are also some cases which can not strictly be counted as errors, but different writing styles. Unfortunately, our performance measure does not account for these, more or less, common cases. Some examples are *pul|dār* “rich”, *doruq|gu* “liar”, and *šaffāf|sāzi* “clarification”. All in all, adding more training samples with the abovementioned features would most probably solve many of the mentioned error groups.

6 Conclusion and Future Work

In this paper, we experimented with different methods to tackle word segmentation correction and zero-width non-joiner recognition problems in Persian. The results on our collected data show BERT outperforming other methods, and the error analysis indicates that it would be relatively easy to increase the performance and pave the way for a practical and effective preprocessing tool. Future work could focus on collecting more informal Persian data, more diverse topics, and more modern registers of the language, to further improve this work’s results. There are also some techniques in the work mentioned in §2, e.g., multi-task learning (say, for the model to learn the difference between joiner and non-joiner characters as an auxiliary task), that could be used in Persian as well. Covering different writing styles in the training data would also be helpful, e.g., changing some of the ending *e[ye]* clitics in Bijankhan corpus, when the word is tagged as GEN, to *ey[e]*, as discussed in §5.

References

- Academy of Persian Language and Literature. 2015. *Persian Orthography*. Academy of Persian Language and Literature.
- Wirote Aroonmanakun. 2002. Collocation and thai word segmentation. In *Proceedings of the 5th SNLP & 5th Oriental COCOSDA Workshop*, pages 68–75.
- Mahmood Bijankhan, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. 2011. Lessons from building a persian written corpus: Peykare. *Language resources and evaluation*, 45(2):143–164.
- Haris Bin Zia, Agha Ali Raza, and Awais Athar. 2018. Urdu word segmentation using conditional random fields (CRFs). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2562–2569, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.
- Ba-Long Bui, Thi-Trang Nguyen, Huu-Hoang Nguyen, and Kiem-Hieu Nguyen. 2019. Hmms for unsupervised vietnamese word segmentation. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6. IEEE.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Nadir Durrani and Sarmad Hussain. 2010. Urdu word segmentation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 528–536, Los Angeles, California, June. Association for Computational Linguistics.
- Schubert Foo and Hui Li. 2004. Chinese word segmentation and its effect on information retrieval. *Information processing & management*, 40(1):161–190.
- Ryo Fujii, Ryo Domoto, and Daichi Mochihashi. 2017. Nonparametric bayesian semi-supervised word segmentation. *Transactions of the Association for Computational Linguistics*, 5:179–189.
- Nguyen Thi Minh Huyen, Azim Roussanaly, Hô Tuong Vinh, et al. 2008. A hybrid approach to word segmentation of vietnamese texts. In *International Conference on Language and Automata Theory and Applications*, pages 240–249. Springer.
- Zhen Ke, Liang Shi, Erli Meng, Bin Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Unified multi-criteria chinese word segmentation with bert. *arXiv preprint arXiv:2004.05808*.
- Sadiq Nawaz Khan, Khairullah Khan, Asfandyar Khan, Asfandyar Khan, Aman Ullah Khan, and Burhan Ullah. 2018a. Urdu word segmentation using machine learning approaches. *International Journal of Advanced Computer Science and Applications*, 9(6):193–200.
- Sadiq Nawaz Khan, Khairullah Khan, and Wahab Khan. 2018b. Supervised urdu word segmentation model based on pos information. *EAI Endorsed Trans. Scalable Information Systems*, 5(19):e2.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Mikhail Korobov. 2015. sklearn-crfsuite.
- John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- Gurpreet Lehal. 2010. A word segmentation system for handling space omission problem in Urdu script. In *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, pages 43–50, Beijing, China, August. Coling 2010 Organizing Committee.
- Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is word segmentation necessary for deep learning of chinese representations? *arXiv preprint arXiv:1905.05526*.
- Chen Lyu, Yue Zhang, and Donghong Ji. 2016. Joint word segmentation, pos-tagging and syntactic chunking. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- Ji Ma, Kuzman Ganchev, and David Weiss. 2018. State-of-the-art chinese word segmentation with bi-lstms. *arXiv preprint arXiv:1808.06511*.

- Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics.
- Ahmad Saffari Moghaddam. 2007. Punctuation in persian. 4:123–137.
- Salar Mohtaj, Behnam Roshanfekar, Atefeh Zafarian, and Habibollah Asghari. 2018. Parsivar: A language processing toolkit for persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Tetsuji Nakagawa and Kiyotaka Uchimoto. 2007. A hybrid approach to word segmentation and pos tagging. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 217–220.
- Cam-Tu Nguyen, Trung-Kien Nguyen, Xuan-Hieu Phan, Minh Le Nguyen, and Quang Thuy Ha. 2006. Vietnamese word segmentation with crfs and svms: An investigation. In *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pages 215–222.
- Dat Quoc Nguyen. 2018. A neural joint model for vietnamese word segmentation, pos tagging and dependency parsing. *arXiv preprint arXiv:1812.11459*.
- Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).
- Mahnaz Panahandeh and Shirin Ghanbari. 2019. Correction of spaces in persian sentences for tokenization. In *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI)*, pages 670–674. IEEE.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. *arXiv preprint arXiv:1603.00786*.
- Dang Duc Pham, Giang Binh Tran, and Son Bao Pham. 2009. A hybrid approach to vietnamese word segmentation using part of speech tags. In *2009 International Conference on Knowledge and Systems Engineering*, pages 154–161. IEEE.
- R. Rashid and S. Latif. 2012. A dictionary based urdu word segmentation using maximum matching algorithm for space omission problem. In *2012 International Conference on Asian Language Processing*, pages 101–104.
- Kashif Riaz. 2010. Rule-based named entity recognition in urdu. In *Proceedings of the 2010 named entities workshop*, pages 126–135. Association for Computational Linguistics.
- Zahra Sarabi, Hooman Mahyar, and Mojgan Farhoodi. 2013. Parsipardaz: Persian language processing toolkit. In *ICCKE 2013*, pages 73–79. IEEE.
- Manabu Sassano. 2002. An empirical study of active learning with support vector machines for japanese word segmentation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 505–512. Association for Computational Linguistics.
- Suteera Seeha, Ivan Bilan, Liliana Mamani Sanchez, Johannes Huber, Michael Matuschek, and Hinrich Schütze. 2020. Thailmcut: Unsupervised pretraining for thai word segmentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6947–6957.
- Mehrnoush Shamsfard, Soheila Kiani, and Yaseer Shahedi. 2009. Step-1: standard text preparation for persian language. In *Third Workshop on Computational Approaches to Arabic Script-based Languages*, pages 859–865.
- Virongrong Tesprasit, Paisarn Charoenpornasawat, and Virach Sornlertlamvanich. 2003. Learning phrase break detection in thai text-to-speech. In *Eighth European Conference on Speech Communication and Technology*.
- Xiaobin Wang, Deng Cai, Linlin Li, Guangwei Xu, Hai Zhao, and Luo Si. 2019. Unsupervised learning helps supervised neural word segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7200–7207.

- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- Jia Xu, Richard Zens, and Hermann Ney. 2004. Do we need chinese word segmentation for statistical machine translation? In *Proceedings of the third SIGHAN workshop on Chinese language processing*, pages 122–128.
- Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. A graph-based model for joint chinese word segmentation and dependency parsing. *Transactions of the Association for Computational Linguistics*, 8:78–92.
- Fan Yang and Paul Vozila. 2014. Semi-supervised chinese word segmentation using partial-label learning with conditional random fields. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–98.
- Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Transition-based neural word segmentation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 421–431.
- Yanna Zhang, Jinan Xu, Guoyi Miao, Yufeng Chen, and Yujie Zhang. 2018. Improving neural chinese word segmentation using unlabeled data. In *IOP Conference Series: Materials Science and Engineering*, volume 435, page 012032. IOP Publishing.