PerSpellData: An Exhaustive Parallel Spell Dataset For Persian

Romina Oji
University of Tehran
Tehran, Iran
romina.oji@ut.ac.ir

Nasrin Taghizadeh
University of Tehran
Tehran, Iran
nsr.taghizadeh@ut.ac.ir

Heshaam Faili
University of Tehran
Tehran, Iran
hfaili@ut.ac.ir

Abstract

This paper presents PerSpellData, a comprehensive parallel dataset developed for the task of spell checking in Persian. Misspelled sentences together with their correct form are produced using a large clean Persian corpus in addition to a massive confusion matrix, which is gathered from many sources. This dataset contains natural mistakes that Persian writers may make which are gathered from a well-known Persian spell checker, Virastman, in addition to the synthetic errors based on a large-scale dictionary. Both non-word and real-word errors are collected in the dataset. As far as we are concerned, this is the largest parallel dataset in Persian which can be used for training spell checker models that need parallel data or just sentences with errors. This dataset contains about 6.4M parallel sentences. About 3.8M is non-word errors, and the rest are real-word errors.

1 Introduction

Every day mass of texts is written with the aid of computers, smartphones, and wearable devices. During typing these texts, several noises are produced because of the writer's fast speed in typing, the lack of knowledge about the correct orthography, or small screens and keyboards on smartphones. Documents with errors are hard to read and even not valuable. Although human reading is robust against misspellings, more time is required to read a misspelled text (Rayner et al., 2006). Therefore, there is a high need for a tool that detects the errors and even corrects them automatically. Spell checkers play an essential role in many applications such as messaging platforms, search engines, etc. (Jayanthi et al., 2020).

A wide variety of spelling correction tools have been created and used in many languages. A top-rated spell checker tool is Grammarly[1]. In Persian, some spell checkers tools were developed such as Virastman[2] and Paknevis[3]. Spelling errors are classified into two categories: non-word and real-word errors (Jurafsky and Martin, 2016). Persian spell checkers detect error words based on a lexicon, so a word is detected as incorrect if it is not in the lexicon. These tools correct errors by using n-grams or a simple shallow neural network model for real-word errors. The most significant disadvantage of these tools is that they do not correct non-word errors within a large context; they show some suggestion words based on window size. Because of the small size of the window, these tools usually cannot correct non-word errors well.

Recent researches on spell checkers in languages such as English show the usefulness of encoder-decoder neural networks for detecting and correcting both non-word and real-word errors (Park et al., 2020; Lertpiya et al., 2020). In general, spell checkers can be considered as a Neural Machine Translation that the incorrect text is in a language and the correct text is the translation in another language. Neural spell checkers that use encoder-decoder models need a large amount of parallel data, therefore, they are usually data-hungry, especially for low resources languages such as Persian. Since there is no publicly available dataset for Persian, the need for a parallel dataset that contains both non-word and real-word errors is of crucial significance. Also, there is no dataset for actual or synthetic real-word errors in Persian.

In this paper, we present the process of making a large-scale dataset for the task of spell checking in Persian. Most of the available Persian datasets were made synthetically (Faili et al., 2016; Mirzababaei et al., 2013; Dastgheib et al., 2019). However, our

[1] https://app.grammarly.com/
[2] http://virastman.ir/
[3] https://paknevis.ir/

dataset, PerSpellData, contains both synthetic and actual mistakes in word and sentence levels. The actual mistakes are collected from two sources: native author's errors and Persian language learner's errors. These data are gathered from Virastman logs and Corpus of Persian Grammatical Errors (CPG)[4].

Shortly, the contributions of this paper can be summarized as follows:

- We present a dataset, PerSpellData, that contains about 6.4M parallel sentences from both formal and informal texts with diverse topics.
- PerSpellData contains both non-word and real-word errors. These errors are actual mistakes humans had made, in addition to the potential synthetic errors. Both word-level and sentence-level errors are covered.
- Synthetic errors are made considering all situations that an error can occur in Persian. These errors are more frequently made by Persian writers.
- The most frequent error type in Persian is word boundary. Specifically, the word به/to is concatenated to the next word.

We made the dataset of about 6.4 million sentences publicly available[5].

The rest of this paper is organized as follows. Section 2 presents the background of work. Section 3 covers an overview of the related works. Section 4 describes the process of making our dataset. Experiments are presented in Section 5. Finally, conclusion and future works are drawn in Section 6.

2 Background

Spelling errors can be categorized into non-word and real-word errors (Jurafsky and Martin, 2016). Non-word errors are the result of a spelling error where the word is not in the lexicon and doesn't have any meaning (like elepant for elephant). Real-word errors are misspelled words when a user mistakenly chooses another word. Real-word errors are valid words but have wrong meaning in their context, or they make the sentence grammatically incorrect (like three are some animals, instead of there).

A confusion matrix is a set of paired words that the first one is a correct word and the second one is the wrong form of the first one. Pairs of confusion matrix show those strings may mistakenly be replaced with each other, like 'there' and 'their' in English. The confusion matrix is the main element of many spell checkers.

3 Related Work

Different strategies used to generate datasets for the task of spell checking can be categorized as follows: 1) generating frequent synthetic errors that writers make (Ahmadzade and Malekzadeh, 2021), 2) generating errors based on features of the language (Bravo-Candel et al., 2021; Bhowmick et al., 2020), 3) gathering errors from human mistakes (Jayanthi et al., 2020), 4) generating errors based on sound similarity (Li et al., 2018), and 5) generating real-word errors based on the similarity of the words in a vocabulary list.

There are several researches on gathering datasets that contain actual mistakes writers made. WikEd Error Corpus (Grundkiewicz and Junczys-Dowmunt, 2014) was automatically extracted from edited sentences of Wikipedia revisions. It was utilized for some enhances in the performance of GEC systems. WikiAtomic Edits (Faruqui et al., 2018) is another dataset that was gathered from Wikipedia Revisions. This corpus contains atomic insertions and deletions of eight languages. GitHub Typo Corpus (Hagiwara and Mita, 2019) is a large-scale dataset of grammatical and spelling errors. It was collected by tracking changes in Git commit histories and gathering typos and grammatical errors. In this dataset, the edits were annotated by native speakers of three languages (English, Chinese, Japanese), and errors were categorized into four categories: mechanical (errors in punctuation and Capitalization), spell, grammatical and semantic (different meaning in source and target).

Some researchers generated synthetic datasets by noising sentences to make parallel misspelled-correct sentence pairs. NeuSpell (Jayanthi et al., 2020) is a toolkit for spelling correction in English, comprising different neural models trained on a syntactic dataset. For each sentence, 20 percent of its words were noised. For injecting error words, character level noise was made randomly or existing confusion matrices were utilized such as

[4] https://ece.ut.ac.ir/documents/76687411/0/CPG.zip

[5] https://github.com/rominaoji/PerSpellData

Table 1: Examples of real-word and non-word errors in English and Persian

Error Type		English Errors		Persian Errors	
		Correct Form	Wrong Form	Correct Form	Wrong Form
non-word	insertion	This story is embracing	This storey is embracing	خوشبختانه همه هنوز دچار نشده اند	خوشبخنانه همه هنوز دچار نشده اند
	deletion	She is an actress	She is an acress	مردم آن شهر خیلی خسته بودند	مردم آن شهر خیی خسته بودند
	substitution	Tehran is the capital of Iran	Tehran is the capitol of Iran	ساعت هفت بیدار می‌شوم	صاعت هفت بیدار می‌شوم
	transposition	He is afraid of bears	He is afraid of bares	از آن‌جا تاکسی گرفتیم	از آن‌جا تاکسی گرتفیم
real-word	insertion	Good jobs are found in big cities	Good jobs are found ink big cities	در این مکان اسکان کنید	در این مکان استکان کنید
	deletion	They live on their own	They live on their on	این مغازه فروشی است	این مغازه فرشی است
	substitution	I cannot see you	I cannot sea you	شلیل میوه خوشمزه‌ای است	دلیل میوه خوشمزه‌ای است
	transposition	I live here	I live heer	این عدد بر مبنای دو است	این عدد بر مبانی دو است
	same pronunciation	This is too much money	This is two much money	این میوه پرتقال است	این میوه پرتغال است
	word boundary	You can do it	Youcan do it	به خانه می‌روم	به خانه میروم

Norvig[6], Wikipedia[7], aspell[8], etc.

In Persian, several datasets were gathered. Corpus of Persian Grammatical Errors (CPG)[9] contains about 700 exam papers of Persian language learners. Dastgheib et al. (2019) used abstracts of Persian papers of various topics and generated a dictionary of correct words. They generated a confusion matrix for this dictionary using Damerau-Levenshtein edit distance (Levenshtein et al., 1966) and sound similarity. They used string distance metric of Kashefi et al. (2013) to find pair of words who differ in one character, which are neighbour in Persian keyboard.

Vafa (Faili et al., 2016) is Persian spell checker that detects and corrects spelling, grammatical and real-word errors. For spelling errors, a confusion matrix was constructed in which the correct words were gathered from Dehkhoda lexicon (Dehkhoda, 1998), and top frequent words of two famous newspaper corpora. Error words are those with 1) one Damerau-Levenshtein distance away for error types of deletion and addition, or 2) two Damerau-Levenshtein distance away for error types of substitution and transposition. Making words noisy was performed regarding some features of Persian; for example, the most frequent characters that may be deleted, or characters that are typed by different hands and may be transposed. In addition to Vafa, another research on Persian real-word errors (Mirzababaei et al., 2013) also used Damerau-Levenshtein distances to generate a confusion matrix.

[6] http://norvig.com/ngrams/spell-errors.txt

[7] https://www.dcs.bbk.ac.uk/~ROGER/wikipedia.dat

[8] https://www.dcs.bbk.ac.uk/~ROGER/aspell.dat

[9] http://search.ricest.ac.ir/dl/search/defaultta.aspx?DTC=36&DC=232735

Table 2: Statistics of PerSpellData.

Errors	Confusion Matrix	PerSpellData
non-word errors	650K	3.8M
real-word errors	1.5M	2.5M
Total	2.15M	6.4M

4 PerSpellData

In this section, we present the process of making PerSpellData, a parallel dataset of misspelled sentences together with the corrected sentences, to improve task of spell checking in Persian. This dataset covers real-word errors and non-word errors. Both of these errors take place because of four kinds of typing mistakes called insertion, deletion, substitution, and transposition. Some Persian and English non-word and real-word errors are shown in Table 1.

Our approach is based on a large corpus of Persian texts in addition to the confusion matrix.

We gathered a confusion matrix containing 2 million pairs of words from various sources, which are explained below. Given the confusion matrix, we made our parallel dataset by replacing correct words in the sentences of corpus with words confusing with them. Table 2 shows some statistics of PerSpellData.

4.1 Corpus and Lexicon

In the first step, we gathered a large-scale Persian corpus. We aggregated three corpora: two of them are CPG[9] and COPER[10], which are publicly available. The third one is corpus of Virastman spell checker, which is about 50 Gigabytes. It is gathered by crawling different Persian Wikipedia pages, articles written in blogfa[11], and news websites like KhabarOnline[12], FardaNews[13], Hamshahry[14], etc. Also, this dataset is cleaned by using autocorrection rules of Virastman.

At the next step, several pre-processing functions were applied on the text in order to clean raw corpus, including normalization of Persian and English characters and numbers, converting symbols to the equivalent text, converting numeric-formatted dates to equivalent text, removing emoji and useless symbols. We used PerSpeechNorm methods for normalization and sentence split (Oji et al., 2021).

All words that appearing in the clean corpus make our lexicon. To ensure the correctness of lexicon words, several annotators checked them manually. Sentences with misspelled words are removed from corpus. Finally, a lexicon with about 290K words is obtained.

4.2 Non-Word Errors

We collected parallel sentences with non-word errors, or confusion matrix to be used to make parallel sentences, from several sources, which are explained below.

Virastman's log: The first and most important source of non-word errors is Virastman's logs. These logs are actual mistakes that users made. There are two cases: 1) user corrected the wrong word by selecting a word among a list of close words that Virastman suggested to the him/her, 2) user corrected the wrong word by replacing with another word rather than the suggested list of Virastman. Virastman logged these two cases and we use them.

Table 3: Different kinds of non-word errors of Virastman log.

Error type	Count	Percentage (%)
word-boundary with space	164,091	53.99
word-boundary with half-space	21,588	7.1
deletion of "ه" and space	12,930	4.25
Replace of "آ" with "ا"	8,513	2.8

Table 4: Distribution of non-word errors of Visrastar log regarding the edit distance between the incorrect word to its correction.

Edit Distance	Count	Percentage(%)
1	234,616	77.2
2	67,999	22.37
3	1,239	0.4
Total	303,903	100

Table 3 presents different kinds of non-word errors extracted from Virastman's logs. About 61 percent of all errors is related to the word boundaries. The distribution of all non-words of Virastman's logs in terms of the edit distance to the correct word is represented in Table 4.

CPG We converted non-word errors of CPG, which is a collection of errors made by Persian learners, to parallel sentences by replacing correct and incorrect forms of errors in the sentences.

FAspell FAspell dataset is a confusion matrix containing Persian spelling mistakes and their correct forms (QasemiZadeh et al., 2006). FAspell has three different error categories: 1) insertion, deletion, substitution, 2) word-boundary, and 3) complex errors, which are mixed of other errors. This confusion matrix was collected from two different sources: first, mistakes made by elementary school students and professional typists; second, wrong words collected from the output of a Persian OCR system. We used only first one, because the second one is very noisy.

Preposition "به/to" A common mistake in Persian writing is related to the preposition "به" when it is concatenated to the next word by mistake and "ه" is also omitted. We manually collected about 500 cases. Some of them are shown in Table 5.

[10] https://github.com/Ledengary/COPER
[11] http://www.blogfa.com/
[12] https://www.khabaronline.ir/
[13] https://www.fardanews.com/
[14] https://www.hamshahrionline.ir/

Close words Close words are those words which are one or two edit-distance away from each other, and one of them has very low frequency in Virastman Corpus, while the other word has a very high frequency. The word with low frequency is not in Virastman Dictionary.

4.3 Real-Word Errors

We gathered real-word errors from different sources, which are explained below.

Virastman's log: Real-word errors that Virastman already has detected as errors and what users selected as correct words make a confusion matrix contains about 1K pair words.

Synthetic confusion matrix: We use Virastman's dictionary of Persian words to make a confusion matrix. This dictionary contains about 290K words. For each word in this dictionary, we find all candidate words that with one or two Levenestain edit-distance (Levenshtein et al., 1966). Therefore, about 1.4 million paired words are created. These errors belong to different categories of insertion, deletion, substitution, transposition, and word-boundary errors.

Informal plural words that use plural signs in wrong ways Some words in Persian stem from Arabic, and they are already plural, but Persian writers wrongly add some plural signs to make these words plural again. We have gathered a list of common plural words in addition to all incorrect forms of them.

Common mistakes in Persian: There are some words in Persian that a wrong form of their writing is common among people. We find these words and the correct form of them from various sources such as Virastaran[15] (a company whose mission is to teach people how to write Persian correctly).

Same sound words: Some words have identical pronunciation but different writing forms. We collect these words using Persian Soundex[16].

Gozar words: There are two verbs in Persian, گزار and گذار, which have the same pronunciation but two different writing styles. Making mistakes in using these two happens because these two words use two different z characters, "ز" and "ذ". Selecting the correct one depends on the word just before them. Sometimes It is even hard for Persian native speakers to select which form is correct. We have gathered about 300 pairs of words which are usually used before them.

CPG dataset: Similar to non-word errors, we converted real-word errors of CPG to parallel sentences by replacing misspelling words with the correct forms.

Tanvin Some Persian words which are rooted in Arabic, have equivalent forms in Persian. We prepared a list of about 100 words containing these words and their correct format. Another issue with Tanvin is that some Persian words must contain it, but writers omit them wrongly, so we have gathered most of these words and their correct forms.

Hamza Two Persian characters, Alef and Yeh, have two different forms of writing (with or without Hamza above), just one of them is correct in each word. Sometimes it is confusing for Persian writers to decide which one is correct. This happens in English too. For example, the word "naïve" can be written as "naive", but the first format is better.

Some examples of the above cases are shown in Table 5.

5 Experiment

To evaluate PerSpellData, we employed a part of this dataset, which is derived from Virastman non-word data logs, containing 1.5M parallel sentences, as the training data and FAspell data with 1600 sentences as the test data. We trained a nested RNN proposed by Li et al. (2018) using NeuSpell implementation[17], referred by CHAR-LSTM-LSTM. In this model, word representations are built by passing individual characters to a char-level bi-LSTM network (CharRNN). Then these representations are passed to a word-level bi-LSTM (WordRNN). The CharRNN collects orthographic information by reading each word as a sequence of letters. The WordRNN predicts the correct words by combining the orthographic information with the context. The hyper-parameters are the same as the original implementation.

The results were compared with Virastman. This tool detects errors using a dictionary and suggests the words using a bi-gram language model and weighted edit distance. Virastman shows related

[15] https://virastaran.net/

[16] https://github.com/feyzollahi/PersianSoundex

[17] https://github.com/neuspell/neuspell

Table 5: Examples of real-word errors in Persian.

Error Type	Example 1		Example 2	
	Correct form	Wrong form	Correct form	Wrong form
Preposition "به"	به‌ویژه	بویژه	به همراه	بهمراه
Make informal plural again plural	اخلاق- خوی‌ها	اخلاق‌ها	عملیات- اعمال	عملیات‌ها
Common mistakes	لپ‌تاپ	لب‌تاب- لپ‌تاب- لب‌تاپ	کارخانه‌ها	کارخانجات
Close words	پزشکی	پرشکی	بهینه	بعینه
Same sound	خواستن	خاستن	قالب	غالب
Gozar words	سپاس گزار	سپاس گذار	گشت‌وگذار	گشت‌وگزار
Tanvin	به ناچار- ناگزیر	ناچاراً	روی هم رفته	اجماعاً
Hamzeh	رئیس	ریییس	متأسفانه	متاسفانه

Table 6: Evaluation of different spell checkers.

Model	Accuracy	Correction Rate
Virastman (all suggestions)	97.95	74.26
CHAR-LSTM-LSTM (Persian)	95.83	58.42
CHAR-LSTM-LSTM (English)	96.60	77.30

suggestions, but it does not perform well on ranking suggestions because it is an interactive spell correction software. Therefore, to evaluate Virastman, all suggestions are considered.

As shown in Table 6, Virastman has high accuracy. It rarely converts correct words to non-correct, so it has a good performance in detecting errors. The accuracy of CHAR-LSTM-LSTM in Persian is higher than in English, because of an extensive dictionary. However, the correction rate is not very good because of the ambiguity of Persian. In Persian, for an incorrect word, there are multiple suggestions that are just one edit distance away. Therefore, it is hard to predict which one is correct. In conclusion, employing a contextualized representation can improve the correction rate of models in Persian.

6 Conclusion and Future Works

In this paper, we presented PerSpellData, which is a parallel dataset for the task of spell checking. We gathered a large scale corpus of Persian text and a confusion matrix of 2 million pairs of words. As the future works, this dataset can be used to train deep encoder-decoder networks to detect and correct both non-word and real-word errors.

References

Ahmad Ahmadzade and Saber Malekzadeh. 2021. Spell correction for azerbaijani language using deep neural networks. *arXiv preprint arXiv:2102.03218*.

Rajat Subhra Bhowmick, Isha Ganguli, and Jaya Sil. 2020. Introduction and correction of bengali-hindi noise in large word vocabulary using rnn. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 277–281. IEEE.

Daniel Bravo-Candel, Jésica López-Hernández, José Antonio García-Díaz, Fernando Molina-Molina, and Francisco García-Sánchez. 2021. Automatic correction of real-word errors in spanish clinical texts. *Sensors*, 21(9):2893.

Mohammad Bagher Dastgheib, SM Fakhrahmad, et al. 2019. Design and implementation of persian spelling detection and correction system based on semantic. *Signal and Data Processing*, 16(3):128–117.

Ali Akbar Dehkhoda. 1998. Dehkhoda dictionary. *Tehran: Tehran University*, 1377.

Heshaam Faili, Nava Ehsan, Mortaza Montazery, and Mohammad Taher Pilehvar. 2016. Vafa spell-checker for detecting spelling, grammatical, and real-word errors of persian language. *Digital Scholarship in the Humanities*, 31(1):95–117.

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. Wikiatomicedits: A multilingual corpus of wikipedia edits for modeling language and discourse. *arXiv preprint arXiv:1808.09422*.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 2014. The WikEd Error Corpus: A corpus of corrective wikipedia edits and its application to grammatical error correction. In *International Conference on Natural Language Processing*, pages 478–490. Springer.

Masato Hagiwara and Masato Mita. 2019. Github typo corpus: A large-scale multilingual dataset of misspellings and grammatical errors. *arXiv preprint arXiv:1911.12893*.

Sai Muralidhar Jayanthi, Danish Pruthi, and Graham Neubig. 2020. Neuspell: A neural spelling correction toolkit. *arXiv preprint arXiv:2010.11085*.

Daniel Jurafsky and James H Martin. 2016. Spelling correction and the noisy channel. *Draft of November*, 7:2016.

Omid Kashefi, Mohsen Sharifi, and Behrooz Minaie. 2013. A novel string distance metric for ranking persian respelling suggestions. *Natural Language Engineering*, 19(2):259–284.

Anuruth Lertpiya, Tawunrat Chalothorn, and Ekapol Chuangsuwanich. 2020. Thai spelling correction and word normalization on social text using a two-stage pipeline with neural contextual attention. *IEEE Access*, 8:133403–133419.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Hao Li, Yang Wang, Xinyu Liu, Zhichao Sheng, and Si Wei. 2018. Spelling error correction using a nested rnn model and pseudo training data. *arXiv preprint arXiv:1811.00238*.

Behzad Mirzababaei, Heshaam Faili, and Nava Ehsan. 2013. Discourse-aware statistical machine translation as a context-sensitive spell checker. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 475–482.

Romina Oji, Seyedeh Fatemeh Razavi, Sajjad Abdi Dehsorkh, Alireza Hariri, Hadi Asheri, and Reshad Hosseini. 2021. Perspeechnorm: A persian toolkit for speech processing normalization.

Chanjun Park, Kuekyeng Kim, YeongWook Yang, Minho Kang, and Heuiseok Lim. 2020. Neural spelling correction: translating incorrect sentences to correct sentences for multimedia. *Multimedia Tools and Applications*, pages 1–18.

Behrang QasemiZadeh, Ali Ilkhani, and Amir Ganjeii. 2006. Adaptive language independent spell checking using intelligent traverse on a tree. In *2006 IEEE Conference on Cybernetics and Intelligent Systems*, pages 1–6. IEEE.

Keith Rayner, Sarah J White, and SP Liversedge. 2006. Raeding wrods with jubmled lettres: There is a cost.