The Impact of Text Normalization on Multiword Expressions Discovery in Persian

Katarzyna Marszalek-Kowalewska
YUKKA Lab AG / Berlin, Germany
k.marszalek.kowalewska@gmail.com

Abstract

This paper evaluates normalization procedures of Persian text for a downstream NLP task - multiword expressions (MWEs) discovery. We discuss the challenges the Persian language poses for NLP and evaluate open-source tools that try to address these difficulties. The best-performing tool is later used in the main task - MWEs discovery. In order to discover MWEs, we use association measures and a subpart of the MirasText corpus. The results show that an F-score is 26% higher in the case of normalized input data.

1 Introduction

The field of computational linguistics (CL) and its engineering domain of natural language processing (NLP) has exploded in recent years. It seems to continue to gain momentum because of a straightforward reason: human civilization is drowning in data. In 2008, Google reported that the Web had one trillion pages.[1] In 2016, the number was estimated to 130 trillions.[2] International Data Corporation projects that by 2025, available data may expand to 175 zettabytes.[3] Although these estimates include video, image data, and databases, most of it is plain old text. Unstructured data (also known as free-form text) comprises 70% - 80% of the data available on computer networks. The information content of this resource is unavailable to authorities, businesses, and individuals unless humans read these texts or devise some other means to derive information value from them. And this is where Natural Language Processing comes to the game. NLP procedures can be applied to characterize, interpret, or understand the information content of a free-form text, in other words, to unlock the potential of unstructured data.

However, since the quality of the input data influences the quality of the output, in most cases before the NLP pipeline uses it, this unstructured data needs to undergo certain cleaning and normalization tasks, e.g., removal of extra whitespace, substitution of acronyms, transformation of numerical information, accent removal, substitution of special characters and emoji, or normalization of date format.

This paper addresses the problem of normalizing texts in the Persian language, which is 5th content language for the Web according to W3Tech.[4] In particular, we will focus on the impact of Persian text normalization on one of the downstream NLP tasks - multiword expressions discovery. MWEs are very frequent in language and it has been proved that their proper treatment can make a significant impact on a number of other NLP tasks, e.g. lexicography (Church and Hanks, 1990; Fellbaum, 2016), word sense disambiguation (Finlayson and Kulkarni, 2011), part-of-speech tagging and parsing (Baldwin et al., 2004), information retrieval (Newman et al., 2012), language learning (Christiansen and Arnon, 2017), machine translation (Carpuat and Diab, 2010) or sentiment analysis (Berend, 2011; Williams et al., 2015). To the best of our knowledge, there have been no previous attempts to analyze the text normalization impact on the discovery of MWEs in the Persian language.

The following section provides a brief overview of previous work on text normalization problem. Then, the specific challenges that the Persian language poses for the NLP tasks are described. Section 4 focuses on the impact normalization has

[1] https://www.itpro.co.uk/604911/google-says-the-web-hits-a-trillion-pages

[2] https://medium.com/@MichelKiflen/google-has-indexed-130-trillion-pages-how-would-you-find-the-one-you-need-d0afa303d6f6

[3] https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html

[4] As of June 2021.

Proceedings of Recent Advances in Natural Language Processing, pages 929–939
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_106

on the discovery of multiword expressions in Persian. It first presents a comparison of different normalization tools and then evaluates the impact of normalizing input data on multiword expressions discovery task.

2 Related Work

Text normalization focuses on transforming noisy (non-standard, informal) text to a more standard representation. Linguistic resources, especially online ones containing slang expressions, acronyms, abbreviations, hashtags, or spelling errors, can deviate a lot from the standard language. Text normalization procedures are applied in order to facilitate NLP applications while dealing with such noisy input.

One of the first studies to indicate the importance of text normalization was done by Sproat et al. (2001) who tried to develop a general normalization process applicable to diverse domains. Since then, the impact of normalizing noisy text and its influence on downstream NLP tasks has been analyzed in a few studies. Han et al. (2013) showed the impact of normalizing social media texts on part-of-speech-tagging. In particular, they focused on tweets and compared original and normalized input texts and different taggers: general Stanford POS tagger and domain-specific Twitter POS tagger. The influence of normalization on parsing was studied by Zhang et al. (2013) who introduced a normalization framework designed with the possibility of domain adaptation. Hassan and Menezes (2013) proposed domain and language independent system based on unsupervised learning for machine translation.

Since text normalization is, in many cases, a necessary preprocessing step for numerous NLP tasks, there are several normalization steps. However, as noticed by Baldwin and Li (2015), it is essential to remember that different normalization tasks would fit different data and downstream NLP applications. Moreover, normalization systems as "one size fits all" seem to be less precise than the tailored ones.

There have been various tasks that consider text normalization as a crucial preprocessing step, the most popular to be spelling correction (Choudhury et al., 2007), statistical machine translation (Aw et al., 2006; Pennell and Liu, 2011) and speech recognition (Kobus et al., 2008). Some unsupervised studies focused on using probabilistic models (Cook and Stevenson, 2009), normalization dictionaries (Gouws et al., 2011), lexicon-based classifiers (Han and Baldwin, 2011) or word association graphs (Sönmez and Özgür, 2014).

Recently, there has also been an interest in applying deep learning for normalization procedures. Baldwin et al. (2015) described several systems taking part in shared tasks of Twitter lexical normalization and named entity recognition, underlining that deep learning systems based on lexicon-augmented conditional random fields (CRFs) achieved the best results. Furthermore, a hybrid neural model, which uses word-based encoder-decoder architecture and a character-level sequence-to-sequence model, was introduced for social media text normalization by Lourentzou et al. (2019). Mansfield et al. (2019) addressed, on the other hand, text normalization problem by directly normalizing full sentences using subword models.

There have also been studies into normalizing less standard or low-resource languages. The impact of normalization was evaluated by Agić et al. (2016) in a study on multilingual projection for parsing low-resource languages. An attempt to normalize dialectal Finnish into the normative standard language was presented by Partanen et al. (2019). Hegazi et al. (2021) studied preprocessing of Arabic text on social media. Research on preprocessing tools, including normalizer for Ainu language, was conducted by Nowakowski et al. (2019). Normalization of six low-resource African languages (Afrikaans, Amharic, Hausa, Igbo, Malagasy, Somali, Swahili, and Zulu) was presented by Zupon et al. (2021).

Research on normalizing Persian text focused mainly on addressing the specific challenges (described in section 3) this language poses for NLP tasks. This resulted in a number of processing tools. In 2010, Shamsfard et al. (2010b) proposed STeP1, which provides tokenization, morphological analysis, part-of-speech tagging, and spell checking. ParsiPardaz toolkit, which, apart from providing the same processing steps as STeP1, also includes additional normalization step, was proposed by Sarabi et al. (2013). The first open-source preprocessing tool - Hazm - was introduced by Hazm (2014). Finally, in 2018 Parsivar, another open-source tool, was presented by Mohtaj et al. (2018). Apart from work on preprocessing tools, research in Persian normalization focused also on classification tree and support vector machine (Moattar et al., 2006), N-gram language model combined

with a rule-based method (Panahandeh and Ghanbari, 2019) or sequence labeling models (Doostmohammadi et al., 2020).

3 Challenges of Persian NLP

Research in Persian NLP faces two significant challenges. The first one arises from the number of resources. Although there has been a significant improvement in the number of available NLP resources e.g. *Hamshahri Corpus* (Darrudi et al., 2004), *Bijankhan Corpus* (Bijankhan et al., 2011), *FarsName* (Hajitabar et al., 2017), *ShEMO* (Nezami et al., 2019), *Persian Dependency Treebank* (Rasooli et al., 2013), *SentiPers* (Hosseini et al., 2018), *FarsNet* (Shamsfard et al., 2010a) or *PersBERT* (Farahani et al., 2020) in recent years, Persian is still a heavily unresourced language compared to English or German. The second problem is related to the challenging character of Persian itself and inconsistencies in the writing system. The following section discusses the main challenges the Persian language poses to NLP applications.

3.1 Encoding

One of the first problems in processing Persian texts is the existence of different character encodings. While creating digital texts, both Persian Unicode characters and Arabic ones are sometimes used. As a result, for example the letter ی [ye] can be expressed by 3 different encodings: either the Persian one: \uo6a9, or two Arabic encodings: \uo6cc or \uo49 (Sarabi et al., 2013; Ghayoomi and Momtazi, 2009; Megerdoomian, 2018).

3.2 Writing System

The Persian writing system poses several difficulties with regard to NLP. First of all, Persian letters can have joiner and non-joiner forms based on their position in a word. This feature is quite common among languages, yet in Persian certain letters written at the end of a word may not be joined. Some users treat them as separate characters and do not use whitespace after the word. As a result, tokenization is not always reliable.

Moreover, foreign (borrowed) elements in Persian tend to be written arbitrarily, i.e., the fact that there are, for example, four possible forms of letter 'z' (ز ظ ذ ض) poses certain difficulties for users. Although the *Academy of Language and Literature*[5] tries to systemize it, there is still great arbitrariness when it comes to actual usage. As an example, consider the following variants of the borrowed word *bulldozer* in Persian:

- بولدوزر
- بولدوظر
- بولدوذر
- بولدوضر

Furthermore, there are no capital letters, which may cause ambiguity for the named entity recognition task. The lack of capital letters can also cause problems with the identification of acronyms.

Another challenge of the writing system is text directionality. Although letters are written from right to left, numbers are written in the opposite direction, e.g.

ایران ۱.۲ میلیون بشکه نفت خام صادر کرد

'Iran exported 1.2 million barrels of crude oil'.

What is more, it is not uncommon for users to use Arabic numerals instead of Persian ones, e.g.

کنفرانس در سال 1997 اتفاق افتاد

'The conference took place in 1997'.

3.3 Word and Phrasal Boundaries

In Persian, as in many other languages, whitespace designates the word boundary. However, apart from the standard whitespace, there is also zero-width-non-joiner space (known as pseudospace) used with non-joiner letter forms. In fact, the whitespace and pseudospace are used inconsistently, causing tokenization and segmentation sometimes really challenging.

As mentioned in 3.2, Persian letters have different forms depending on their position in a word. Thus, users often treat non-joiner forms incorrectly, i.e., not adding whitespace after them, e.g., توازماگرفت 'YouTookFromUs'. As a result, this phrase would be processed as one lexeme instead of four separate ones, i.e., تو از ما گرفت.

On the other hand, whitespace is often used instead of pseudospace which causes words such as زبان‌شناسی 'linguistics' to be processed as two separate words زبان 'language' and شناس 'knowledge'

[5] Academy of Language and Literature (In Persian فرهنگستان زبان و ادب فارسی) is the official Iranian regulatory body of the Persian language.

(when written with whitespace, i.e., زبان شناسی). As a result, word and phrase boundaries are often unclear, and tokenization, phrase segmentation, and clause splitting can be very challenging steps in the Persian NLP pipeline.

Inconsistent use of white- and pseudospace is directly related to complex lexemes, consisting of a lexeme and attached affixes that represent a separate lexical category or part of speech from the one they are attached to. A few examples of this situation are presented in table 1.

3.4 Ambiguity

Dealing with word sense ambiguity is one of the main NLP challenges. This task is particularly difficult in the case of Persian as the number of heterophonic homographs (words with identical written forms but with different pronunciations, each associated with a different meaning) is high. The main reason for this situation is the fact that Persian short vowels are usually not written. Therefore, the word ملک could be interpreted in the four following ways:

- ملک [malak] 'angel',
- ملک [malek] 'prince',
- ملک [melk] 'domain',
- ملک [molk] 'country, territory'.

3.5 Ezafe Construction

Ezafe is a syntactic construction used to express determination. In most cases, it is pronounced but not written (since it is expressed by a short vowel), contributing to ambiguity, especially in chunking and semantic as well as syntactic processing of a sentence. Hence, the following sentence can be interpreted in two different ways depending on the presence of ezafe:

پدر حسن را دید

1. [pedar hasan ra did] 'Father saw Hassan.'
2. [pedare-e hasand ra did] 'He/She saw Hassan's father.'

4 Normalization Impact on Multiword Expressions Discovery

The challenges presented above: inconsistency in using white- and pseudospace, different encodings, missing short vowels or bidirectionality can pose many difficulties for proper processing of Persian for several NLP tasks. Therefore, a certain level of text normalization seems necessary. The following section describes the impact normalization procedures have on the discovery of multiword expressions task.

4.1 Multiword Expressions Discovery

Linguistics expressions that consist of at least two words (even when represented by a single token) and are syntactically and/or semantically idiosyncratic - this is probably the most common definition of multiword expressions. They attracted a lot of research attention and have been the main topic in plenty of papers.

MWEs are very frequent in language and range over a number of different linguistic constructions, from idioms, e.g. *to kick the bucket*, to fixed expressions, e.g. *fish and chips*, light verb constructions, e.g. *give a demo*, to noun compounds, e.g. *traffic light*. Biber et al. (1999) claim that the number of MWEs in spoken English is 30% – 45% and 21% in academic prose. Jackendoff (1997) suggests that the number of MWEs in a speaker's lexicon is the same as simple words, yet if we take into consideration the domain-specific lexicons, this number seems to be an underestimation (Sag et al., 2002). Indeed, the research conducted by Ramisch (2009) suggests that the MWEs ratio can be between 50% and 80% in a corpus of scientific biomedical abstracts. Research by Krieger and Finatto (2004) estimate that MWEs can constitute more than 70% of the specialized lexicon.

MWEs processing consists of two tasks: identification and discovery (Constant et al., 2017). MWEs identification focuses on tagging a corpus with actual MWEs. The research on MWEs in Persian has so far focused mainly on the identification of verbal multiword units and light verb constructions (LVCs) in particular, e.g., Taslimipoor et al. (2012); Salehi et al. (2012, 2016). MWE discovery - the task this paper tries to address - is a process that focuses on finding new MWEs (types) in corpora and storing them, e.g., in the form of a lexicon, for further usage. This task takes text as input and generates a list of MWE candidates from it. These candidates can be further filtered and evaluated by trained experts. True MWEs are stored in a repository or added to the MWE lexicon. To our knowledge, there have not been any studies that address the discovery of MWEs in Persian with

Word	Type	Whitespace	Pseudospace	Attached
به	Preposition	به شیوه	به‌شیوه	بشیوه
هم	Prefix	هم کلاس	هم‌کلاس	همکلاس
این	Determiner	این مرد	این‌مرد	اینمرد
آن	Determiner	آن قدر	آن‌قدر	آنقدر
را	Postposition	شرایط را	شرایطرا	شرایطرا
که	Relativizer	چنان که	چنان‌که	چنانکه

Table 1: Complex tokens (Ghayoomi and Momtazi, 2009).

respect to the normalization of input text.

The assumption that MWEs stand out, i.e., they exhibit some sailence, allows us to extract (or discover) them automatically from texts. This salience is also why especially statistical procedures, such as association measures (AMs), have been so popular when it comes to MWEs discovery. This paper also approaches the discovery of MWEs by employing a selected set of association measures.

4.2 Corpus

The corpus used in the study was MirasText (Sabeti et al., 2018) corpus - an automatically generated text corpus for Persian. It is one of the largest available Persian corpora, containing 2.8 million documents and over 1.4 billion tokens. The corpus size is 15GB. Each data point is provided with the following information: content, title, content summary and keywords, base website, and exact URL of the webpage.

The content of the MirasText corpus was generated from 250 web pages selected from a wide range of fields to ensure the diversity of data, e.g., news, economy, technology, sport, entertainment, or science.

Corpus content was generated through crawling; thus, there is a possibility of including duplicated texts. In order to remove duplicated content from the corpus Sabeti et al. (2018) used a filtering process based on a bloom filter (Almeida et al., 2007).

4.3 Normalization

4.3.1 Processing Tools Evaluation

To ensure that the best normalization tool is used for the discovery of MWEs task, firstly, research comparing two open-source processing tools for Persian was carried out. These tools are Hazm and Parsivar, and they both provide normalization, tokenization, chunking, and part-of-speech steps. In order to evaluate these tools, a small corpus of 5000 sentences was annotated by 3 Persian linguistic experts with respect to sentence segmentation and tokenization. The inter-annotator agreement was calculated with *Fleiss' Kappa* - a metric used to evaluate the agreement between three or more raters (Fleiss, 1971) and annotators achieved 98% which indicates *almost perfect agreement*. [6] Table 2 presents the tokenization results of normalized and raw data.

	Precision	Recall	F1
Hazm not-normalized	71%	73%	72%
Hazm normalized	97,5%	97%	97%
Parsivar not-normalized	79%	75%	77%
Parsivar normalized	99%	98%	98%

Table 2: Tokenization results.

The better tokenizer turned out to be Parsivar (Mohtaj et al., 2018) achieving 98% F-score. The superior performance of Parsivar over Hazm was also confirmed in the Persian plagiarism detection study by (Mohtaj et al., 2018). It seems that the main difference between these two tools lies in the better performance of space correction by Parsivar.

Nevertheless, what seems to be of higher importance here is the fact that the results obtained using raw and normalized corpus differ significantly. Regardless of the preprocessing tool used, the tokenizer performance was in both cases more than 20% higher in the case of the normalized data.

4.3.2 Corpus and Its Normalization

Since the MirasText corpus data was obtained via crawling, it seems necessary to perform certain cleaning and normalization tasks. The initial corpus analysis showed that a certain number of articles contain incomplete content (clipped content). Such articles were excluded from the final corpus used in this study. After filtering out the clipped articles, the total number of corpus documents was 2,072,521. As the next step, 50 million token corpus for the discovery of MWEs was sampled.

For most of the NLP tasks, the first necessary step is to tokenize the input text. However, as already mentioned, this is not a simple task in Persian text processing since there are two kinds of spaces:

[6] For interpretation see Landis and Koch (1977).

white- and pseudospace, which are not used consistently. Using inconsistent spacing results in high ambiguity, both on lexical and syntactical levels. Therefore, for a corpus of millions of documents written by thousands of various authors, it is necessary to unify its data, and one of the first and most essential unification steps in Persian NLP is to correct spaces.

As a result of an experiment described in 4.3.1, the best processing tool turned out to be Parsivar (Mohtaj et al., 2018), and the corpus used for the discovery of MWEs task was normalized with it. Parsivar, in its normalization task apart from encodings and numbers unification, performs two different types of space correction:

- rule-based space correction: a set of rules using regular expressions were employed in order to detect spaces within words correctly, e.g., می روم (miravam) 'I am going' or تحلیل گر (tahlilgar) 'analyzer'. The problem with words that consist of two or more tokens but cannot be extracted with one of these rules was addressed by constructing a dictionary. This helped with words as گفت و گو (goft-e gu) 'conversation'.

- learning-based space correction: using training model that recognizes multi-token words as one token. Parsivar uses 90% of the Bijankhan corpus (which contains multi-word tokens annotated with IOB tagging format) as training data. Naïve Bayes model was used to find word boundaries. The model was evaluated on the remaining 10% of Bijankhan corpus and got 96.5% of F-score for space correction on that validation set.

Table 3 presents raw and normalized metrics of sentence segmentation and tokenization performed on the corpus used in the present study.

Task	Not-normalized	Normalized
number of sentences	1,464,996	1,537,725
number of tokens	52,536,988	51,525,867

Table 3: Evaluation metrics of the corpus.

As can be seen, both the number for sentence segmentation and tokenization differ significantly (the difference in the number of tokens is almost 1 million!) if we compare the corpus before and after normalization. The difference in sentence segmentation stems from the incorrect treatment of dots in the not-normalized corpus, especially in case of numerals, dates, webpages and in combination with other punctuation marks. These results show that proper cleaning and normalization tasks (especially unifying spaces) are crucial during Persian text processing.

4.4 Methodology

In order to extract Persian multiword expressions, a list of 20 lemmas that would serve as initial seeds was prepared. The task of MWEs discovery was addressed from a statistical perspective. For every lemma, its bi-grams and tri-grams were extracted separately from raw and normalized corpus using the following association methods:

- PMI
- log-likelihood
- t-score
- χ^2 test

These particular AMs were chosen as they are the most popular ones used for the discovery of MWEs (Evert, 2008; Seretan, 2008; Wahl and Gries, 2018; Villavicencio and Idiart, 2019).

For each association measure, its top 100 bi- and tri-grams per lemma were extracted - this resulted in 1487 unique MWE candidates from the normalized corpus and 1817 from the raw one.

5 Results

5.1 Candidates Filtering

The outcome of employing association measures to discover Persian MWEs is a list with 1487 unique MWE candidates from the normalized corpus and a list with 1817 unique MWE candidates from the raw corpus. All MWE candidates were evaluated by trained experts - Persian native speakers with linguistic background.

Annotators were provided detailed guidelines which included an operational definition of MWEs ("Multiword expressions (MWEs) are lexical items that: a) can be decomposed into multiple lexemes and b) display lexical, syntactic, pragmatic and/or statistical idiomaticity" as presented by Sag et al., 2002) and a number of examples presenting true and false MWEs. Each MWE candidate was evaluated by at least three annotators who answered the question: Is the following sequence a valid multiword expression? Possible answers include: YES, NO, and UNABLE TO DETERMINE.

The total number of experts contributing to this project was 21, and the inter-annotator agreement (IAA) was calculated again with *Fleiss' Kappa*. All annotators were working on both sets: MWE candidates extracted from raw and normalized corpus. The IAA results were 87% and 81% for normalized and raw corpus, respectively. Thus, the final average IAA for this task was 84% which indicate that *almost perfect agreement* was achieved.

5.2 Multiword Expression Discovery Evaluation

After evaluating the candidates, the number of true MWEs in a normalized corpus was 389 and 154 in the raw one.

The main objective of this study was to evaluate the impact of text normalization on the MWEs discovery task in Persian. Figure **??** shows the performance of the four selected association measures when it comes to the discovery of true MWEs.

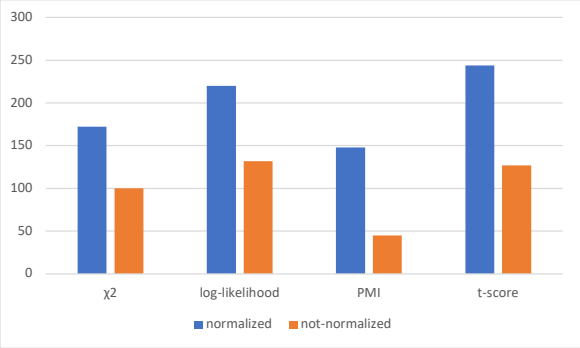

Figure 1: A number of true MWEs extracted with analyzed association measures.

As can be seen, each AM performs better when used with the normalized data. The highest number of MWEs were extracted with t-score (248 MWEs), followed closely by log-likelihood (220 MWEs), both performed on the normalized corpus.

The number of true MWEs is, however, not enough to evaluate the performance. Therefore, it is interesting to perform error analysis and check which cases were and which were not discovered in the raw corpus (compared to the normalized one). Correctly detected MWEs in the raw corpus can be divided into three categories:

- MWEs with Arabic numerals, e.g. 360 پانوراما *360 panorama*,
- MWEs with words written in Latin script, e.g., HDR فناوری *HDR technology*,
- MWEs whose components do not contain non-joiner letters, e.g., ابر کامپیوتر *supercomputer*.

True MWEs discovered in the normalized corpus but not in the raw one seem to have generally one thing in common: they contain words with non-joner letters; therefore, the use of whitespace is not always consistent. Examples of MWEs discovered in normalized corpus but missed in the raw one are فروش برخط *online sales*, بازی رایانه‌ای *computer game*, باشگاه ورزشی *sport club*, or اکولوژی دریا *marine ecology*. Furthermore, all MWEs found in the raw corpus were also discovered in the normalized one.

In order to further evaluate true MWEs discovered using raw and normalized corpus, we used the combined outcome from all AMs. For MWE candidates from raw and normalized corpus, precision, recall, and F-score were computed (similarly to Evert and Krenn, 2001 who used these metrics to plot a precision-recall curve for direct comparison of different AMs). The overall impact of text normalization on the discovery of multiword expressions in Persian is presented using F-score in table 4.

	F-score
Not-normalized	15%
Normalized	41%

Table 4: Comparison of F-score.

The F-score turned out to be 26% higher in the case of normalized data. Therefore, applying text normalization procedures proved to have a significant impact on the discovery of multiword expressions task in Persian.

Since the different normalization steps may vary in the impact on the downstream NLP tasks, their performance for discovering MWEs in Persian was also analyzed. Table 5 presents F-score for all text normalization steps (performed separately as well as in various combinations). It turned out that the most efficient combination of normalization steps is the unification of encodings and dates combined with space correction. In fact, correcting and unifying spaces proved to be the most crucial normalization step for the presented task.

6 Conclusion and Future Works

In this paper, an impact of text normalization on a downstream NLP task was presented. In particular, we focused on the normalization of Persian language data for multiword expression discovery.

Normalization step(s)	F-score
Encodings unification	19%
Date unification	21%
Space correction	35%
Pinglish conversion	18%
Encodings unitication + date unification	24%
Encodings unification + space correction	39%
Encodings unification + date unification + space correction	**41%**
Encodings unification + date unification + space correction + pinglish conversion	40%

Table 5: Evaluation of normalization steps on MWEs discovery in Persian.

The experiment results show that the performance of a system without a Persian-tailored normalization step is 26% worse (F-score), which is a significant deterioration. To our knowledge, this was the first time when the influence of text normalization on the discovery of multiword expressions in Persian was described.

Since this paper focuses on normalization as a preprocessing step, it would be interesting to compare its impact with post-processing tasks. Some further future works include analyzing how normalized data influences other NLP tasks in the Persian language, particularly syntactic parsing and sentiment analysis. Moreover, we would like to compare the tools described in this paper with a neural network approach to text normalization.

Acknowledgments

The author would like to thank the anonymous reviewers for their encouraging feedback and insights.

References

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Paulo Almeida, Carlos Baquero Sérgio, Nuno Preguiça, and David Hutchison. 2007. Scalable bloom filters. *Information Processing Letters*, 101:255–261.

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for SMS text normalization. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 33–40, Sydney, Australia. Association for Computational Linguistics.

Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China. Association for Computational Linguistics.

Tyler Baldwin and Yunyao Li. 2015. An in-depth analysis of the effect of text normalization in social media. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 420–429, Denver, Colorado. Association for Computational Linguistics.

Gábor Berend. 2011. Opinion expression mining by exploiting keyphrase extraction. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1162–1170, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education Ltd., Essex, England.

Mahmood Bijankhan, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. 2011. Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*, 45:143–164.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245, Los Angeles, California. Association for Computational Linguistics.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Basu Anupam.

2007. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition (IJDAR)*, 10(3–4):157–174.

Morten H. Christiansen and Inbal Arnon. 2017. More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, 9:542–551.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Paul Cook and Suzanne Stevenson. 2009. An unsupervised model for text message normalization. In *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 71–78, Boulder, Colorado. Association for Computational Linguistics.

Ehsan Darrudi, Mahmoud R. Hejazi, and Farhad Oroumchian. 2004. Assessment of a modern Farsi corpus. In *Proceedings of the 2nd Workshop on Information Technology and Its Disciplines*, Iran.

Ehsan Doostmohammadi, Minoo Nassajian, and Adel Rahimi. 2020. Joint Persian word segmentation correction and zero-width non-joiner recognition using BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4612–4618, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Stefan Evert. 2008. Corpora and collocations. In A. Lüdeling and M. Kytö, editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter, Berlin.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France. Association for Computational Linguistics.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. ParsBERT: Transformer-based model for Persian language understanding. *ArXiv*, abs/2005.12515.

Christiane Fellbaum. 2016. The treatment of multiword units in lexicography. In Philip Durkin, editor, *The Oxford Handbook of Lexicography*, pages 411–424.

Mark Finlayson and Nidhi Kulkarni. 2011. Detecting multi-word expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 20–24, Portland, Oregon, USA. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Masood Ghayoomi and Saeedeh Momtazi. 2009. Challenges in developing Persian corpora from online resources. In *Proceedings of the 2009 International Conference on Asian Language Processing*, Singapore. IEEE Computer Society.

Stephan Gouws, Dirk Hovy, and Donald Metzler. 2011. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 82–90, Edinburgh, Scotland. Association for Computational Linguistics.

Alireza Hajitabar, Hossein Sameti, Hossein Hadian, and Arash Safari. 2017. Persian large vocabulary name recognition system (FarsName). In *2017 Iranian Conference on Electrical Engineering (ICEE)*, pages 1580–1583.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology*, 4(1):5:1–5:27.

Hany Hassan and Arul Menezes. 2013. Social text normalization using contextual graph random walks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1577–1586, Sofia, Bulgaria. Association for Computational Linguistics.

Hazm. 2014. Python library for digesting Persian text.

Mohamed Osman Hegazi, Yasser Al-Dossari, Abdullah Al-Yahy, Abdulaziz Al-Sumari, and Anwer Hilal. 2021. Preprocessing Arabic text on social media. *Heliyon*, 7(2).

Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. 2018. SentiPers: A sentiment analysis corpus for Persian. In *Computing Research Repository (CoRR)*.

Ray Jackendoff. 1997. Twistin the night away. *Language*, 73:534–559.

Catherine Kobus, François Yvon, and Géraldine Damnati. 2008. Normalizing SMS: are two metaphors better than one ? In *Proceedings of the 22nd International Conference on Computational*

Linguistics (Coling 2008), pages 441–448, Manchester, UK. Coling 2008 Organizing Committee.

Maria Krieger and Maria José Bocorny Finatto. 2004. *Introdução à terminologia: teoria & prática*. Contexto, Sao Paulo.

Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Ismini Lourentzou, Kabir Manghnani, and ChengXiang Zhai. 2019. Adapting sequence to sequence models for text normalization in social media. In *International Conference on Web and Social Media*, pages 201–206. AAAI.

Courtney Mansfield, Ming Sun, Yuzong Liu, Ankur Gandhe, and Björn Hoffmeister. 2019. Neural text normalization with subword units. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 190–196, Minneapolis, Minnesota. Association for Computational Linguistics.

Karine Megerdoomian. 2018. Computational linguistics. In Anousha Sedighi and Pouneh Shabani-Jadidi, editors, *The Oxford Handbook Of Persian Linguistics*, pages 461–480.

Mohammad Hossein Moattar, Mohammad Mehdi Homayounpour, and D. Zabihzadeh. 2006. Persian text normalization using classification tree and support vector machine. In *2006 2nd International Conference on Information and Communication Technologies*, pages 1308–1311. IEEE.

Salar Mohtaj, Behnam Roshanfekr, 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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David Newman, Nagendra Koilada, Jey Han Lau, and Timothy Baldwin. 2012. Bayesian text segmentation for index term identification and keyphrase extraction. In *Proceedings of COLING 2012*, pages 2077–2092, Mumbai, India. The COLING 2012 Organizing Committee.

Omid Mohamad Nezami, Paria Jamshid Lou, and Mansoureh Karami. 2019. ShEMO – a large-scale validated database for Persian speech emotion detection. *Language Resources and Evaluation*, 53:1–16.

Karol Nowakowski, Michal Ptaszynski, Fumito Masui, and Yoshio Momouchi. 2019. Improving basic natural language processing tools for the Ainu language. *Information*, 10(11).

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.

Niko Partanen, Mika Hämäläinen, and Khalid Alnajjar. 2019. Dialect text normalization to normative standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 141–146, Hong Kong, China. Association for Computational Linguistics.

Deana Pennell and Yang Liu. 2011. A character-level machine translation approach for normalization of SMS abbreviations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 974–982, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Carlos Ramisch. 2009. Multi-word terminology extraction for domain-specific documents. Master thesis, Grenoble, France.

Mohammad Sadegh Rasooli, Manouchehr Kouhestani, and Amirsaeid Moloodi. 2013. Development of a Persian syntactic dependency treebank. In *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Behnam Sabeti, Hossein Abedi Firouzjaee, Ali Janalizadeh Choobbasti, S.H.E. Mortazavi Najafabadi, and Amir Vaheb. 2018. MirasText: An automatically generated text corpus for Persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, CICLing-2002, pages 1–15, Mexico City, Mexico.

Bahar Salehi, Narjes Askarian, and Afsaneh Fazly. 2012. Automatic identification of Persian light verb constructions. In *Computational Linguistics and Intelligent Text Processing. CICLing 2012. Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2016. Determining the multiword expression inventory of a surprise language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 471–481, Osaka, Japan. The COLING 2016 Organizing Committee.

Zahra Sarabi, Hooman Mahyar, and Mojgan Farhoodi. 2013. Parsipardaz: Persian language processing toolkit. In *Proceedings of The 3rd International Conference on Computer and Knowledge Engineering, (ICCKE)*, Iran.

Violeta Seretan. 2008. *Collocation extraction based on syntactic parsing*. Ph.d. thesis, University of Geneva.

Mehrnoush Shamsfard, Akbar Hesabi, Hakimeh Fadaei, Niloofar Mansoory, Ali Famian andSomayeh Bagherbeigi, Elham Fekri, Maliheh Monshizadeh, and S. Mostafa Assi. 2010a. Semi automatic development of FarsNet; the Persian WordNet. In *5th Global WordNet Conference*.

Mehrnoush Shamsfard, Hoda Sadat Jafari, and Mahdi Ilbeygi. 2010b. Step-1: A set of fundamental tools for Persian text processing. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Cagil Sönmez and Arzucan Özgür. 2014. A graph-based approach for contextual text normalization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 313–324, Doha, Qatar. Association for Computational Linguistics.

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer Speech and Language*, 15(3):287–333.

Shiva Taslimipoor, Afsaneh Fazly, and Ali Hamze. 2012. Using noun similarity to adapt an acceptability measure for Persian light verb constructions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 670–673, Istanbul, Turkey. European Languages Resources Association (ELRA).

Aline Villavicencio and Marco Idiart. 2019. Discovering multiword expressions. *Natural Language Engineering*, 25(6):715–733.

Alexander Wahl and Stefan Th. Gries. 2018. Multiword expressions: A novel computational approach to their bottom-up statistical extraction. In P. Cantos-Gómez and M. Almela-Sánchez, editors, *Lexical Collocation Analysis: Advances and Applications*, pages 85–109. Springer International Publishing, Cham.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42:7375–7385.

Congle Zhang, Tyler Baldwin, Howard Ho, Benny Kimelfeld, and Yunyao Li. 2013. Adaptive parser-centric text normalization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1159–1168, Sofia, Bulgaria. Association for Computational Linguistics.

Andrew Zupon, Evan Crew, and Sandy Ritchie. 2021. Text normalization for low-resource languages of Africa. In *African NLP. EACL 2021*.