

Challenges and Opportunities in Processing Low Resource Languages: A study on Persian

Mehrnoush Shamsfard

NLP Research Lab, Shahid Beheshti University, Tehran, Iran
m-shams@sbu.ac.ir

Abstract

This paper discusses the importance of language processing and its challenges. It first defines low resource languages and their influencing factors. Then talking about the Persian Language, discuss the situation of Persian in this field of study. Following the discussion, some major Persian language resources are introduced. and describing available methods, some challenges and opportunities are discussed. At last the conclusion section suggests some steps for moving Persian from a low resource language toward a high-resource one, including using cross-lingual embeddings, standardization of test data, running various challenges on Persian data and encouraging startups to build their business in this field.

Keywords: Language technology, Low resource language, Persian

Résumé

این مقاله به اهمیت پردازش زبان طبیعی و چالشهای آن می پردازد. به این منظور ابتدا با ارائه تعریف زبانهای با منابع محدود، عوامل تاثیرگذار در این محدودیت را بررسی می‌نماید. سپس با بحث در مورد زبان فارسی، به جایگاه آن از نظر محدودیت منابع اشاره می‌نماید. در ادامه برخی از منابع زبانی توسعه‌یافته برای زبان فارسی معرفی شده، کمبودهای این منابع مورد توجه قرار می‌گیرند. مقاله سپس با بیان روشهای فعلی، به بحث در مورد چالشها و فرصتهای موجود در پردازش زبان فارسی پرداخته و در نهایت در بخش نتیجه گیری راهکارهایی برای افزایش منابع زبانی آن ارائه می‌نماید.

1. Introduction

There are 6-7 thousand languages in the world with different size of native speaker population. Some of these languages such as English, Spanish or French are very popular to be learnt and spoken as the second language (L2) while some others may be just spoken or written by a small number of native speakers.

Most of the languages which have orthography and written form have written documents and many of written languages have electronic documents on the web or local media. So being able to process these documents is a must. Resource-rich languages are those with matured language technology, many language resources and processing tools and applications and on the opposite side, resource-poor languages suffer from the lack of data, language resources, language technology and processing tools and applications. Actually a small fraction of languages is resource rich or high resource, among which, English has a specific position; it is actually a laboratory language which has attracted a big share of computational linguistic efforts and researches in the world. Even non-native English speakers work on English language to develop or enhance resources, algorithms, methods, processing tools and applications or create data or text to present their ideas, news, achievements and knowledge in this language (this paper is an evidence of this fact). In this way, some few resource rich languages are used more and more in the digital world and cyberspace and the others are weakened. This forces the speakers of the weakened languages to learn those resource-rich languages to be able to transfer their ideas, present their culture and language, talk about their achievements, and so on. Thus the resources of resource rich languages become richer every day and riches get richer.

On the other hand, the weakened languages will gain lower share of documents in the cyberspace and move to being removed from digital media gradually. Paying attention to these languages and providing essential technologies to enable their speakers to communicate and share their data, experiences and knowledge in their own native language not only helps to save diverse languages but also makes the data, information and knowledge from different cultures, geographic regions, political governments, etc. available to linguists, social studying and aiding communities, weather and environment watchers, politicians, militaries and so on. Thus, enhancing language technologies for low resource languages is beneficial to all.

2. Low Resource languages

There are different definitions for low resource languages. Although low-density or indigenous languages are included in this category, it also includes some languages with a large size of native speaker population. LORELEI defines low resource languages as the languages for which no automated human language technology exists. This definition excludes languages for which there are some tools but they do not cover all aspects of human language or does not have good performance such as Persian. According to Duong (2017) a language is considered low-resource for a given task if there is no algorithm using currently available data to automatically do the task with adequate performance. This definition implies that a language is considered low-resource based on a specific task. In this definition for example, Persian is not a low-resource language with respect to part-of-speech tagging of formal written text as the performance of a tagger is 96% precision, while it is low resource for POS-tagging of colloquial informal texts or for other tasks such as recognizing multi word expressions.

There are various factors affecting languages; keeping them low resource or helping them to become high-resource. Among these factors the availability of language resources and technology, expertise in natural language processing, financial supports and political issues are the most important. These factors are themselves effective on each other in a cycle.

The most important factor which makes a language low-resource is the technical one; availability of language resources and technology. In the next section we talk about its challenges and opportunities.

The other factor which is highly influenced by the others is human expert availability. Many researchers from all over the world (even low resource language native speakers) are serving language processing technology for resource-rich languages such as English, because

- There are sufficient available English training data for machine learning algorithms and so novel methods can be tested easily.
- There are many tested, reliable tools and toolkits, and many pieces of source codes or reproducible methods available for conducting a research for English while starting the same research for a low resource language is very hard and time- and effort- consuming and may lead to lower performance measures at the end.
- Journals and conferences related to language technology usually ask authors to compare their work with other similar works and as in many cases there is no similar work in the low resource language the researcher has to pay attention to methods in resource-rich ones such as English too.
- As the audiences of a language in minority are much less than the audiences of languages in majority, publishing research paper about minor languages in journals is harder. Many Journals does not accept such papers due to their audiences' needs and even after publication the paper may have fewer citations. This problem lead some researchers to change their work focus from their own native language to the English language.

Two other factors are economics and politics. Countries with rich economy can dedicate sufficient financial resources and funds for research on language technology while the others cannot. The rich countries may even be interested on processing some non-native languages due to some political, economic or cultural issues so their economy may help other language resources to grow.

On the other hand, political issues may affect or prevent the development of language resources and enhancement of language technology. For example, in Iran which is under sanctions, the sanctions are applied to any aspect of Iranian lives including science and technology. Unavailability of powerful processing devices, source files and libraries from code providers such as google, some softwares even compilers and interpreters, data, and even rejecting research papers due to political issues are some of the problems which sanctions have brought to Iranian researchers in this field.

All of these factors and some others make a language as low resource or help it to grow to high resource level.

In the rest of the paper, we study the situation of Persian in this field of study and propose some steps toward providing language resources and technology for Persian

3. The Persian Language

According to traditional classification, Persian with the Indo-Aryan languages constitutes the Indo-Iranian group within the Satem branch of the Indo-European family. This group consists of Persian, Pashto, and Kurdish.

Persian is the official language of Iran, Afghanistan and Tajikistan with more than one hundred million speakers and also is spoken in more than six other countries. According to its geographical position Persian speaker countries are neighbor to Arabic countries and so there are a lot of loaned words entered to Persian from Arabic. Although there are many differences between Persian and Arabic language especially according to grammar and syntactic features, there are some similarities in lexical level and some Arabic derivational rules have come into this language (Shamsfard, 2011).

Although Persian is the official language of Iran, there are some other languages spoken in Iran such as Kurdish, Turkish and Arabic and sometimes documents to be processed are a mixture of these languages. Even for Persian, there are various types and dialects. For example, Persian texts can be written in colloquial or formal Persian. Colloquial texts are used in daily conversations, non-formal short messages (SMS), blogs, social media, emails and some books (especially novels) while formal texts are used in formal conversations, formal or scientific documents, news, educational and many other books. Persian colloquial and formal texts are very different from each other especially in the lexical level. They need different lexicons, different training and testing datasets, and even different grammatical rules for NLP tasks. Most of the language resources and tools are dedicated to formal Persian.

Some linguistic features of the formal Persian language are as following (Shamsfard, 2011);

Persian is written right-to-left. It is a pro-drop language with canonical SOV word order with a lot of frequent exceptions in word order, which have turned Persian to a free word order language. Verbs are marked for tense and aspect and agree with the subject in person and number with some exceptions. Although verb-final, Persian is otherwise mostly head-initial.

Persian letters have one to four forms of writing. Different forms are used depending on the position of the letter within the word which may be initial, medial or final (isolated). There are various scripts for writing Persian texts, differing in the style of writing words, using or elimination of spaces within/between words, using various forms of characters and so on. Persian is a derivational and generative language in which many new words may be built by concatenating words and affixes. Usually none of the short vowels are written in a Persian sentence. So facing homographs and homonyms are popular ambiguities in Persian. Usually there is no definite article in a Persian sentence while most of the nouns appear with one in English. Unlike English there is no female/male distinction for Persian pronouns and there is no rule for appearing uncountable nouns in singular form. Even words which are uncountable may appear in plural form.

In Persian each verb conjugates in its own tense while in English the tense of the other sentence verb must be considered. In Persian words and phrases may be omitted in a sentence according to a syntactic or semantic symmetry. Omitting the subject is also very popular in

Persian sentences. In this case the agreement (person and number) embedded in the verb can play the subject role.

In Persian in many cases adjectives can be inserted in place of nouns without any lexical change and this may cause structural or semantic ambiguities in noun phrases.

Working on Persian language processing is a growing field. The early efforts in this field go back to late 1980s and the very first systems such as Dena for Persian text understanding (Fahimi & Shamsfard, 1995) were introduced in 1990s. Many language resources, tools and applications have been developed for Persian during the last 25 years. But still Persian is far from English in language technology and can be assumed a low resource language in many tasks. In the next section we mention some of the main language resources for Persian but left the survey on tools and applications for another paper due to the small size of this paper.

4. Language Resources for Persian

Available language resources can be divided into the following categories. In each category some are named.

- Corpora: There are various corpora available for Persian, with different sizes and taggings. Peykareh (Bijankhan et al., 2011) with about 8 million tagged tokens, FLDB (Assi, 1997), and Hamshahri (AleAhmad et al., 2009), are the most famous general POS tagged corpora. The large Beheshti corpus of more than 4 billion tokens, dump of Wikipedia¹, blogfa corpus and corpus of tweeter posts are raw corpora used for language modeling and building word embeddings. PAYMA (Shahshahani, et al., 2015), ArmanPerosNERCorpus (Poostchi, et al., 2016), A'laam (Hosseinnejad, et al., 2017) are instances of NER tagged corpora for Persian. Parallel corpora such as Mizan (Kashefi, 2018) with more than 1 million sentence pairs and TEP (Pilevar, et al., 2011) with about 550,000 pairs of movie subtitles, and comparable corpora such as (Hashemi, et al., 2010) with 7500 document pairs are another type of corpora. Task specific corpora such as Mahtab plagiarism detection corpus (Mashhadirajab, et al., 2016) with 20000 documents and about 10000 suspicious documents, Beheshti sense tagged corpus (Rouhizadeh et al., 2019), Treebanks such as (Mirzaei and Safari, 2018) for discourse and (Rasooli et al., 2013) for dependency fall in this category too.

- Lexicons and Thesauri: General lexicons such as zaya (Eslami, et al., 2004) with 55000 entries and FarsVajeh (Shamsfard and Jafari, 2017) with about 80000 entries are available. In Farsvajeh, each lexeme is associated with its various written forms and the preferred orthography suggested by APLL, its phonetics, POS tags, frequency in a corpus, and morphological structure (inflectional, derivational, compound, ...). Although there are some larger lists of words, we couldn't find lexicons larger than 100,000 entries with phonetic, morphological, and syntactic information. Of course the electronic version of some Dictionaries such as sokhan, dekhoda, mo'een, etc. are available (mostly without a legal license) but they are not structured. There are also some general thesauri such as Fararooy (1998) (almost a translation of Ruget's) and some domain specific thesauri such as the ones developed by Irandoc² (eg (Norouzi, 2003). Sentiment lexicons such as

PerSent (Dashtipour et al., 2016), LexiPers (Sabeti et al., 2016)), HesNegar (Asgarian, et al., 2018) and SentiFars (Dehkharghani, 2019) fall in this category too.

- Wordnets and knowledge graphs: FarsNet is the first, biggest and most reliable wordnet for Persian. The third version of it contains more than 100,000 lexical entries organized in more than 40,000 synsets with glosses, examples and various semantic relations. It is developed semi-automatically and revised manually. Persian wordnet of Tehran (Taghizadeh and Faili, 2018) is another work in this field. It is translated automatically from Princeton WordNet. FerdowsNet as the third Persian wordnet which is smaller than the first two is developed by Ferdowsi University but is not available to public. According to wordnet Persian is among resource-rich languages. FarsBase (Asgari, et al., 2019) is the Persian knowledge graph with 5,582,589 links to external datasets.

- Datasets: the main problem in Persian resources is here. Persian lacks large reliable data sets for training and testing systems for different NLP tasks. Even for tasks that have large known English datasets such as question answering (QA), chatbots, text generation, WSD, multi-word expression (MWE), sentiment analysis, entailments and paraphrases, etc., there is either no dataset or the available datasets are too small or not reliable. Some of the available datasets are wsd data (Rouhizadeh, et al. 2019), Pars-ABSA for aspect based sentiment tagged opinions (Ataei, et al., 2019), and parallel formal and informal Persian language (under construction),

It seems that the open problems are development of corpora with various tags (except POS and NER) and datasets for various application tasks such as those mentioned in the previous paragraph.

5. Methods, Challenges and Opportunities

In recent years, the language technology is shifted from rule based systems to statistical ones and now to deep neural networks and distributed semantics with dense vectors. In the current trend of research, we need a huge amount of data to train deep neural systems and low resource languages are those for which such a data is unavailable. Even for unsupervised methods which do not need training data, at least we need standard, accurate, reliable test datasets with good coverage to test the developed systems. Persian is among the languages which suffer from the shortage of language resources such as tagged corpora and train and test datasets. For instance, datasets of questions and answers, entailment sentences, paraphrase pairs, chats, restyling sentences, metaphors, and texts and their internal representations are some examples which are available for English but not for Persian. While datasets for tasks such as syntax parsing over constituency or dependency (tree banks), sentiment analysis, named entity recognition, word sense disambiguation and machine translation (parallel corpora) are available for Persian but mostly their size, and sometimes their accuracy and quality are lower than the corresponding data in English.

The shortage can be eliminated by either creating resources and tools for the low resource languages from scratch or trying to adapt/use resources and tools in other languages.

¹ <https://dumps.wikimedia.org/fawiki/>

² <https://irandoc.ac.ir>

Thus we can assume four major approaches to handle the shortcoming of data and resources for Persian language:

- Direct Translation of data from English or any other language which has the data: Unfortunately, this approach is not a good one in many cases, as the quality of translators are not admissible and the translated data should be revised and corrected manually which takes a long time itself. The created dataset is biased to the source language and may omit the linguistic phenomenon of Persian. On the other hand, due to differences of the two languages, the word-by-word alignment may be impossible or some features such as POS tag may be changed during the translation and so the method may be inappropriate for some sorts of datasets or some types of tasks (eg. POS tagging). The small dataset for Persian WSD (Rouhizadeh, et al., 2019) is an example of this method.
- Processed translation of English or other language's data: In this method, in addition to translation, the data will be under some processes to enhance the translated data and remove uncertain parts. In other words, this approach utilizes cross-lingual methods and bilingual (or parallel) resources to build Persian datasets from English ones. This method is more complicated than the previous one but its results are more reliable than translation without human revision. It is faster than the next method and can create larger datasets with lower costs. The method is more suitable for translating words and not texts and has the drawback of biasness same as the previous method. The Persian wordnet developed by Tehran University (Taghizadeh, et al, 2018) is an example of this method.
- Creating data for Persian: Many researchers try this approach. It is time and cost consuming and the results are usually smaller than the corresponding datasets in English. But the dataset is not biased to any other language and it's usually more accurate and precise than the previous ones. FarsNet (Shamsfard, et al. 2010; khalghani & Shamsfard, 2018)), Dataset of Mahtab (Mashhadirajab et al., 2016), Persian treebank (Rasooli, et al, 2013) and ArmanPerosNERCorpus (Poostchi, et al., 2016) are some examples of this method.
- Using cross-lingual and transfer based deep methods to use English or other language's data to perform Persian tasks: Moving toward cross-lingual efforts is a way to use e.g. English embeddings for Persian tasks. It seems that for each deep approach to a problem if the embeddings be cross-lingual then the method will work for Persian as well as the English one using English training set. This method is expected to be faster than the previous ones and can utilize larger, more reliable datasets for Persian tasks.

6. Conclusion and Future Work

In some domains we have enough resources while in some others, the resources are either rare or missing. For example, there are enough POS and NER tagged corpora, good wordnets, huge amount of untagged raw texts, several sentiment lexicons, and medium size parallel (Persian-English) corpora. But still, some fundamental tasks such as tokenization have problem in available corpora and datasets. So compound words and verbs are not assumed as one token and this makes a lot of problems in various tasks especially in word embedding.

We need enhanced corpora, tagged by various features such as entity linking and coreferences and various datasets such as training and testing data for question answering (QA), chatbots, text generation, WSD, multi-word expression, sentiment analysis, entailments and paraphrases, etc.

It seems that (1) moving towards Cross-lingual embeddings, (2) establishing a research center for standardization and generation of test data and creating testbeds and benchmarks to evaluate resources and tools and (3) running various challenges on Persian data such as those ran by SemEval, SenseEval, TREC, ... and (4) encouraging and helping startups and companies to build their business in this field, are some possible actions which will speed up the development of Persian resources and move this language from being low-resource.

7. Bibliographical References

- AleAhmad, A. Amiri H., Darrudi E., Rahgozar M., and Oroumchian F. (2009). Hamshahri: A standard Persian text collection. *Knowledge-Based Systems Journal*.
- Asgari B., M., Hadian A., Minaei-Bidgoli M., (2019). FarsBase: The Persian knowledge graph, *Semantic Web*,
- Asgarian E, Kahani M, Sharifi S. HesNegar: Persian Sentiment WordNet (2018). *JSDP*. 15 (1):71-86
- Assi, M. (1997). Farsi Linguistic Database (FLDB), *The International Journal of Lexicography*, 10(3):6, Oxford University Press.
- Ataei, T.S., Darvishi, K., Minaei-Bidgoli, B., Eetemadi S., (2019). Pars-ABSA:an Aspect-based Sentiment Analysis dataset for Persian, arXiv:1908.01815.
- Bijankhan, M. Sheykhzadegan, J. Bahrani, M. and Ghayoomi, M., (2011). "Lessons from Building a Persian Written Corpus: Peykare," *Language Resources and Evaluation*, 45(2):143–164.
- Dashtipour, K., Hussain, A., Zhou, Q., Gelbukh A., Hawalah A.Y.A., Cambria E., (2016). PerSent: A Freely Available Persian Sentiment Lexicon. In BICS 2016, 8th International Conference on Brain-Inspired Cognitive Systems.
- Dehkharghani R., (2019). SentiFars: A Persian Polarity Lexicon for Sentiment Analysis, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2).
- Duong, L., (2017). Natural language processing for resource-poor languages, PhD dissertation, University of Melbourne, Australia.
- Eslami, M., Sharifi, M., Alizadeh, S., Zandi, T., (2004). 'Persian Generative Lexicon', 1st workshop on Persian Language and Computer, pages 6-11.
- Fahimi, M., Shamsfard, M. (1995) Dena: A Persian Text Understanding System, In Computer conference of Computer Society of Iran (CSI).
- Fararooy, J., (1998). *Persian Thesaurus*, Ibex Publishers.
- Hosseini, P., Ahmadian R. A., Maleki, H., Anvari, M., Mirroshandel, S.A., (2018). SentiPers: A Sentiment Analysis Corpus for Persian, arXiv:1801.07737.
- Hosseinnejad S, Shekofteh Y, Emami Azadi T. A'laam Corpus: A Standard Corpus of Named Entity for Persian Language. (2017). *Journal of Signal and Data Processing (JSDP)*. 14 (3):127-142.
- Kashefi O., (2018). MIZAN: A Large Persian-English Parallel Corpus, arXive

- Khalghani, F., Shamsfard, M., (2018). Extraction of Verbal Synsets and Relations for FarsNet. The 9th Global WordNet Conference (GWC 2018).
- Mashhadirajab, F., Shamsfard M., Adelkhah R., Shafiee F., and Saedi C. (2016). A Text Alignment Corpus for Persian Plagiarism Detection. In FIRE 2016, pages 184-189.
- Mirzaei, A. and Safari p., (2018). Persian Discourse Treebank and Coreference corpus, LREC 2018, pages 4049-4055.
- Norouzi, M., (2003) *Engineering Thesaurus*, Irandoc.
- Pilevar M.T., Faili H., Pilevar A.H. (2011) TEP: Tehran English-Persian Parallel Corpus. In Computational Linguistics and Intelligent Text Processing. CICLing 2011. Pages 68-79. Springer, Berlin, Heidelberg.
- Poostchi, H., Borzeshi, E.Z., Abdous, M., Piccardi, M. (2016). PersonER: Persian Named-Entity Recognition. In Proceedings of COLING 2016, pages 3381–3389, Osaka, Japan.
- Rasooli, M.S., Kouhestani, M., Moloodi A., (2013). Development of a Persian Syntactic Dependency Treebank, NAACL, pages 306-314, Atlanta, Georgia, USA.
- Rouhizadeh, H., Shamsfard, M., Rouhizadeh M., (2019). Knowledge based word sense disambiguation with distributional semantic expansion. Widening NLP Workshop, ACL2019.
- Sabeti, B., Hosseini, P., Ghassem-Sani G.R., Mirroshandel, S.A., (2016). LexiPers: An ontology based sentiment lexicon for Persian, In GCAI 2016. 2nd Global Conference on Artificial Intelligence, pages 329-339.
- Shahshahani, M., Mohseni M., Shakery A., Faili H., (2019). PEYMA: A Tagged Corpus for Persian Named Entities, *Journal of Signal and Data Processing (JSDP)* 16 (1) :91-110.
- Shamsfard, M. (2011) Challenges and open problems in Persian text processing. In: 5th Language & Technology Conference (LTC).
- Shamsfard M., Hesabi A., Fadaei H., Mansoory N., Famian A., Bagherbeigi S., Fekri E., Monshizadeh M., and Assi S. M (2010) Semi-automatic development of FarsNet; the Persian wordnet, In Proceedings of 5th global WordNet conference, Mumbai, India.
- Shamsfard, M., Jafari, H.S. (2017) FarsVajeh: The Persian lexicon. Technical report, NLP Research Lab, Shahid Beheshti University, Tehran, Iran.
- Taghizadeh N., Faili H., (2016). Automatic Wordnet Development for Low-resource Languages using Cross-Lingual WSD. *Journal of Artificial Intelligence Research* 56(56):61-87.