Language Art, 6(1): pp.75-90, 2021, Shiraz, Iran
DOI: 10.22046/LA.2021.05 DOR: 98.1000/2476-6526.1399.6.75.18.1.65.110
Article No.: 61.52.139912.7590

ORIGINAL RESEARCH PAPER

A Survey of Part of Speech Tagging of Latin and non-Latin Script Languages: A More Vivid View on Persian

Dr. Meisam Moghadam[1]©
Assistant professor, Faculty of science, Fasa University, Iran.

Niloofar Jafarpour[2]
MA, Institute of Linguistics and Literary Studies
Technical University of Darmstadt, Germany.

(Received: 3 June 2020; Accepted: 28 November 2020; Published: 28 February 2021)

This research is a general overview of the Latin script languages part of speech (POS) tagging with a specific focus on the non-Latin script languages, especially Persian. The study reviews the progress in POS tagging among the 23 highest native spoken languages in the world. Some of these languages follow the right-to-left (RTL) writing system such as Arabic, Urdu and Persian which have their own specific issues in POS tagging. This paper also goes through the issues and challenges which occurs during the tokenization and part of speech tagging of these languages. The challenges can be common between the languages or be specified to one. The Persian Language is chosen as the main interest of this paper and an attempt is made to critically overview the recent studies on Persian part of speech tagging and enumerate the specific challenges occurring in these studies. Reviewing the bulk of literature and examining the features, challenges, issues, and POS tagging tools in Persian, it was concluded that significant challenges of the researches on Persian were generally in the tokenization level and mostly as a result of using the Arabic script and its characteristics.

Keywords: Part of Speech Tagging, Latin Script Language, Non-Latin Script language, RTL system, Persian Language.

.

[1] E-mail: moghaddam.m@fasau.ac.ir © (Corresponding Author)

[2] E-mail: niloofar.jafarpour@yahoo.com

Introduction

Part of speech tagging is an essential task in the natural language processing. Different approaches have been applied to the Latin script languages such as English and German and to the East Asian languages.

Comparatively, less researches and studies have been done on the South Asian languages or the other languages written in the different scripts, although these langauges have considerably high range of native speakers all over the world, as reported by Khanam (2013).

In this paper, it is tried to go through only the 23 highest native spoken languages and investigate their issues and challenges in the part of speech tagging procedure.

If we compare the 23 languages with at least 50 million first-language speakers in the world concerning their progress in NLP and particularly part of speech tagging, a point comes over relating to the little progress of this branch of study in these languages.

The list below is taken from the online active research project of world's known living languages, ethnologue.com, which shows the 23 languages of the world with the highest native speakers. These languages are, Chinese, Spanish, English, Arabic, Hindi, Portuguese, Bengali, Russian, Japanese, Lahnda, Javanese, Korean, German, French, Telugu, Marathi, Turkish, Urdu, Vietnamese, Tamil, Italian, Persian and Malay.

In the following parts of this paper, first, an overview of Latin scripts languages is given. Afterwards, it is tried to investigate the non-Latin script languages and mention some of the features of these languages, which can lead to specific challenges in part of speech tagging. Finally and mainly a closer investigation will be done on Persian part of speech tagging.

Part of speech tagging of Persian Language is a young research area, but the reported results are significantly successful.

Part of speech tagging of Latin script languages

Probably when Dionysius Thorax of Alexandria was working about 2115 years ago on his grammatical sketch of Greek and the list of eight part of speeches, he himself could not expect that his work will have such a remarkable proportion of the Greek, Latin and most European language basis till recent modern linguistics era. However, nowadays, the popular tag-sets for English consist of more lexical classes, which are mostly derived from 87 tags of the Brown corpus. (Jurafsky and Martin, 2009).

Part of speech tagging of English has a rich background and different approaches have been developed successfully and being used by other researchers working on the other languages.

The Transformation and Discourse Analysis Project (Harris, 1962), dealing with the ambiguities with 14 hand-written rules, can be mentioned as the earliest implementation of part of speech algorithm which is re-implemented by Joshi and Hoely (1999) and Karttunen (1999), as mentioned by Jurafsky and Martin (2009).

The Rule-based or stochastic algorithm are mostly two approaches leading the part of speech tagging methods. As Seraji (2015) noted, the EngCG, based on Constraint Grammar architecture (Karlsson et al, 1995) can be an example of rule-based tagger.

One example of stochastic algorithm is Hidden Markov Model (HMM). Brill (1995) believed that the combination of these two approaches makes the transformation based taggers.

Between the years 1990 and 2000 the attention and tendency to data-driven taggers have been increased. Trigrams'n'Tags (Brants, 2000) can be mentioned as an example.

Tokenization as a non-trivial step in the language processing is to some extent simpler for inflectional languages like English as the space is used as the word barrier. Although even in English or such space delimited languages, there are still some complications in dealing with multi-word expressions.

Part of speech tagging for non-Latin script and RTL (right-to-left) writing system Languages

In this part, the discussion centers on the part of speech tagging of non-Latin Script languages including Arabic, Indian langauges, Urdu, Asian langauges and Chinese. It is possible to put forward the claim that the similarities between the languages bring similar challenges. To deal with unsolved challenges of different languages, it is really helpful if one has a general view on the features of different languages and finds common issues between the languages.

Arabic

Arabic Script writing system which is also used by non-Arabic speaking states, is the third most widely used writing system after Latin and Chinese. Recently, the Arabic natural language processing have been developed increasingly as its importance of being spoken by nearly 500 million people around the world. As Khoja (2001) mentioned, it is considerable to mention that there are syntactical, morphological and semantical differences between Arabic and Indo- European languages. In Arabic language, a fixed pattern and certain infix, prefix and suffix are being used on the roots for conjugation. The role of the root in forming verbs and most of the nouns is a significant characteristic in Semitic languages and also plays an important role in the stemming process. However, stemming and assigning the tags have their own issues due to the complexity of the conjugation. An example of that is the converting of some letters to the other letters in case of adding some

affixes or the ambiguities of some letters which can be part of the word or affixes. (Khoja, 2001)

The significant difference between Arabic and Indo-European languages puts forward the need of separate tag set for this language. The part of speeches of Arabic are categorized into noun, verb and particles. Analyzing the different Arabic-Script languages show that they have many challenges in common including, no short vowels, no capitalization in Arabic-script, and huge number of ambiguities (Arabic, 19.2) or the presence of variant forms of Arabic-script in the other language texts which are mostly the challenges appearing in tokenization process.

Indian Languages

The Indian languages including Hindi, Bengali, Telugu, Marati and Tamil are rich morphological languages written mostly in Devanagari script from left to right.

The first efforts on part of speech tagging of these languages was rule based approach. The need of considerable knowledge about the language to assign the written rules brings these approach into challenge. The appliance of these approach to the Indian languages brings the need of a big tag set which makes tagging the parts of speech difficult. Moreover, the high existence of ambiguities in these languages makes it a hard task to assign part of speech tag of words according to the text it involves. As mentioned before, the rich morphology of Indian Language makes new issues in part of speech tagging. An instance arising in comparison of Hindi language to English language indicates that while English has 7 to 8 inflected word forms, in Hindi this number can reach 40 forms. (Vikram, 2013)

The existing developed part of speech tagger are in Hindi, Bengali, Panjabi and Tamil Languages which are mostly based on statistical and Hybrid approaches.

Urdu

Urdu language, like Persian language, is an Indo-European Language. Accordoing to Khanam (2013), Urdu is written in Arabic script (Persoarabic) though it is not a Semitic language. This language is spoken by 150 million people around the world as native or second languages. The morphology of Urdu language has high influence of Arabic, Persian, Turkish, Sanskrit and Hindi Languages. In this language, there is no delimiter in most of hand written texts or the delimiter is inconsistent. Like Persian alphabet, Urdu alphabet consists of joiners and non-joiners alphabet but the alphabets have significant differences, (Rehman et al., 2013) In Arabic and Persian languages, the omission of delimiters applies to specific rules and does not happen so frequently even in the hand written texts.

Asian Languages and Chinese

The tokenization process of some Asian Languages such as Thai, Lao or Chinese, which are without systematically mark word in a text, has its own issues. (Jurafsky and Martin, 2009) As an example of these group of languages, in Thai

language the segmentation can be based on the longest matching technique. Basically, the algorithm starts reading the text, looking for the longest match in the dictionary. In case of finding the match but not receiving the allowance to find the rest, the algorithm starts again looking for another match. (Rehman et al., 2013)

Chinese language with the highest native speaker than other languages, mostly spoken in China Taiwan, Singapore and Malaysia, is written in symbols and has no alphabet.

The first challenge that one faces in the Chinese language part of speech tagging is the absence of word boundaries in this language. Accordingly, segmentation, either before or simultaneous with part of speech tagging is required. (Tou Ng et al., 2004)

Another challenge, according to Xia, (2000) would be the lack or little inflectional morphology existing in Chinese language which is normally expected for assigning the part of speech tags. (Xia, 2000)

Xia's (2000) findings lend support to the claim that tagging criteria according to the syntactic distribution of the word would be a better choice than basing only on the meaning in Chinese language because of its complying with the theories of contemporary linguistics.

POS tagging in Persian Language

Considering Behistun Inscription of the Achaemenid Darius I, the old Persian dates back to more than 3000 years ago 522 - 486 BCE. However, what can be considered of morphology of the language dates back to middle Persian, the Persian spoken at the era of the Parthian Empire (248 BCE - 226 CE) and the Persian during the Sassanid Empire (226 - 651 CE). Middle Persian Grammar is too similar to what we know as modern Persian Grammar except the difference in vocabulary which is the result of Arab invasion on Persia. (Curtis and Tallis, 2005) The first attempt in part-of-speech tagging of Persian language (Assi & Abdolhoseini, 2000), which followed the method of Schuetze (1995), reported the accuracy of 69-83% for the numbers and different types of verbs and nouns and generally 57.5% for the automatic part of the system. The resulted accuracy for adjective and adverbs was really low and using this method could not disambiguate part of speech of the words and the less frequently used words in the text. In this tagger a tag-set with 45 tags was used.

Brants (2000) introduced Orumchian tagger for Persian POS tagging which follows the TNT POS tagger. The TNT tagger is based on Hidden Markov Models theory. This system uses 2.5 million tagged words as training data and the size of the tag-set is 38. Reported accuracy of this approach is 96.64% reported.

Another research for Persian POS tagging is done by Megerdoomian (2004). Explaining some of the linguistically challenges in the development of Persian POS tagging with no experimental result is the only result of this research.

In another research, Raja et al. (2007) used TNT tagger for Persian language, reporting the general accuracy of 96.64%, specifically, 97.01% on the known words and 77.77% on the unknown words. This accuracy was better than the reported accuracy of Spanish language and close to English and German Language. The tag-set used in this tagger had 38 tags. (Raja et al., 2007)

Raja et al. (2007) presented evaluation of some tagging methods on texts in old version of Peykare (Textual Corpus of the Persian Language). By ignoring many morphosyntactic features of words, the number of tags in the tag set decreases to 40. The Raja (2007) tagger, based on the Memory and Maximum likelihood Approach, training the tagger on 85% of it and letting 15% to be tested, resulted similar performance to the other languages such as English, Spanish and German. The important point in this work was the experimentation of simple heuristics that could be applied in post- processing of the output of the tags, which had a positive impact on the improvement of tagging of unknown words especially for the weaker models. This Heuristic was basically a modification of a few prefix or suffix characters of the word which was tested giving to the level of the post-processing of the tags. (Raja et al., 2007)

Another approach was what Azimizadeh, Arab and Quchani (2008), based on Finite-State –Transducer (FST), used. This part-of-speech tagger was a part of a Persian text-to-speech system, Pars Gooyan. The final system accuracy had the result of 83.51%.

The Azimizadeh et al, (2008) tagger was based on Hidden Markov Model (HMM), included in Pars Gooyan System, which was an implementation in festival TTS software, reporting overall results of 95.11%, 96.136% for the known words and 60.25% for the unknown words. (Quchani et al., 2008)

Fadaei and Shamsfard (2008) presented an algorithm to tag Persian unknown words. Using 60 inflectional and derivational affixes and a set of 140 rules, they try to analyse words morphologically. The algorithm detects the probable affixes in the word, constructs and prunes the word's parse tree, calculates the truth probability of the remaining derivations and in the last step it assigns the most probable tags to the words. There are some ambiguities in this work. The number of tags and the tagset are not uttered in the paper. Also used corpus and its details are not described.

Mohtarami (2008) tagger is based on using Heuristic Rules to improve Persian part of speech tagging accuracy. The Maximum Likelihood of Estimation (MLE) approach is used as well in purpose of evaluating the effects of those rules, because

it was simpler to be implemented. This tagger reported the result of 95.29% accuracy. (Mohtarami et al., 2008)

Mohseni and Minaei-bidgoli (2010) described a method based on morphological analysis of words for a Persian Part-Of-Speech (POS) tagging system focusing on Peyekare (or Textual Corpus of Persian Language). Peykare is arranged into two parts, annotated and unannotated parts, while the annotated part is taken into account in order to create an automatic morphological analyzer.

Okhovvat, and Minaei Bidgoli (2011) implemented a part-of-speech tagging system on Persian corpus by using hidden Markov model. To achieve this goal, the main aspects of Persian morphology was introduced and developed. To evaluate the accuracy of their proposed approach, the approach was applied in simulations which were done on both homogeneous and heterogeneous Persian corpus. Getting results with 98.1% accuracy in the experiments demonstrate the suitable efficiency of the proposed approach on Persian corpus.

Forsati and Shamsfard (2012) examined Bees colony algorithm to find the most probable tag for a word. They employed stochastic information as its fitness function. In the same line, Seraji, Megyesi, and Nivre (2012) designed a dependency parser for Persian language and discovered the linguistic dependencies to ease NLP tasks.

Kardan and Imani (2014) used maximum entropy as a classifier for POS tagging. They chose those types of features that can show the most important characteristics of a word. Nourian, Rasooli, Imany, and Faili (2015) used dependency grammar, on Ezafe detection and improved its precision rate. Pakzad and Minaei Bidgoli (2016) also used dependency grammar and joint probability for Persian and English annotation.

Furthermore, Hosseini Pozveh, Monadjemi, and Ahmadi (2016) employed artificial neural networks for POS tagging due to their ability to learn complex patterns. The accuracy rates of 95.7% and 96.17% were reported. Comparing the results with the results obtained from other approaches makes it obvious that neural networks can do POS tagging and named entity recognition more accurately than other methods.

Some of the taggers mentioned above are not open- source part of speech tagger. The reports show good results in using of several POS tagging methods like TnT, Memory based tagger (MBT) and Maximum Likelihood Estimation (MLE). (Raja et al., 2007) Most of the mentioned experiments used Bijankhan Corpus.

Available corpora for Persian language

One of the corpora for Persian language is Bijankhan corpus (Bijankhan, 2004) with 2,597,939 tokens. Moreover, Upsala Persian Corpus (Seraji, 2015) which is an available corpus with 2,704,893 tokens. Some other morphologically annotated

corpora which are not freely available are as the following: The Persian Linguistic Data base (Assi, 2005) with 56 million words from contemporary text, Hamshahri collection (AleAhmad et al., 2009), Cooperative Persian-English Corpus (Hashemi et al., 2010) in which the Persian part is created based on Hamshahri news agency and the English part from BBC news agency, Peykare (Bijankhan et al., 2011) containing 110 million words and Mizan English-Persian Parallel corpus (Mizan, 2013) containing one million English sentences often from classic literature with its translation in Persian. (Seraji, 2015)

Different syntactically annotated corpora exist in Persian language as well. Farsi Linguistic Database (FLDB) corpus by Assi (1997) comprises a selection of contemporary Modern Farsi literature, formal and informal spoken varieties of the language, and a series of dictionary entries and word lists (about 3 million).

Amtrup, Mansouri Rad, Megerdoomian, and Zajac (2000) created Shiraz corpus which is a bilingual tagged corpus developed from a Persian corpus of on-line material to test machine translation project at New Mexico State University.

Taghiyareh, Darrudi, Oroumchian, and Angoshtari (2003) used a text collection that contains laws and regulations passed by Iranian Parliament which is a small-sized collection focusing on one subject category.

Another Persian corpus is Mahak, provided by Sheykh Esmaili, Abolhassani, Neshati, Behrangi, Rostami, and Mohammadi (2007) that is prepared for evaluation of information retrieval systems. Also, this corpus contains 3007 documents.

Challenges in Part of Speech Tagging of Persian Language

As Mohseni, Motalebi, Minaei-bidgoli, Shokrollahi-far (2008) mentioned, according to the different structure of Persian language, there is some challenges which are not seen in some other languages like English. Moreover, as Hosseini Pozveh, Monadjemi, and Ahmadi, (2016) mentioned, Persian language is a free word order language in which the base structure frequently changes and words can place in different positions. This feature can be assumed as a challenge in POS tagging. It is considerable to mention that the Arabic script languages come from different language families which have different morphological rules, challenges and issues. The Persian Language itself, which is a branch of Indo European family, is divided structurally into three categories: Farsi, spoken in Iran; Dari spoken in Afghanistan and Tajiki spoken in Tajikistan and Uzbekistan. Tajiki is written since 1940 in Cyrillic alphabet with some favor between people to switch to Latin alphabet. (Beeman, 2005)

The Arabic script used in Persian brings with itself different challenges in Persian natural language processing.

In modern Persian the three short vowels of the six existing vowels, are shown by diacritics. The designing of the system for part of speech tagging should be

flexible to detect the non-written short vowels in Persian texts. This characteristic of Persian Language causes homographs and ambiguities because some same letters in a word can have different pronunciation and completely different meaning and therefore, different part-of-speech tags. In the Persian literary texts, this characteristic is being used as a powerful literary figure of speech (Seraji, 2015).

مردم

Transcription	meaning
mrdm	Unclear
mærdom	People
mærdæm	I am a man.
mærdæm	my husband
mordæm	I died.

Table 1: An example of ambiguity of Persian homographs of word مردم with and without Diacritics on the letters م and د

Moreover, this characteristic in Persian language makes the Persian noun phrase really ambiguous. Hosseini Pozveh et al. (2016) enumerated ambiguity as another main concern of POS tagging. Ambiguity refers the fact that a word has more than one grammatical role or interpretation. For example, the word "interest" can be a noun or a verb. Persian language also faces the ambiguity problem. They examplified the word /ʃirin/ which can be a proper noun (name) as well as an adjective which means "Sweet". Moreover, due to rapid changes on different sciences, it receives a lot of new words from other languages. Those words mostly fall in the ambiguous class and harden NLP applications.

Although by detecting the pronouns and proper names at the end of a noun phrase or the suffix *را* as an indicator for an object noun phrase or affixes such as pronominal clitic, it is possible to create a tag for marking the boundaries of noun phrases, still without a written linking constituent between the nouns in a noun phrase and few existing overt morphemes to specify noun phrase boundaries, it stays as an ambiguous issue in Persian part of speech tagging (Amtrup et al., 2000).

In Persian language tenses are fewer than English language. This language has wide derivational and inflectional morphology. Persons inflect Verbs and the syntax is not influenced from gender. According to Mohseni et al. (2008), like English language, Derivational Persian words are extracted by prefixing and suffixing their stems. One of these issues is related to the numerous categories of verbs in Persian language with various inflections in relation to persons which lead to variety forms of words.

Mohseni et al. (2008) examplified same forms which can mean various morphemes. For example, the suffix "ی" can be considered as a connecting part for

the second person e.g. "رفتی" singular or as the indefinite piece of a word e.g. "کتابی". This challenge is known as ambiguities in Persian morphology.

As in the other Arabic Script languages, the whitespace between words is an issue in tokenization which leads to different ambiguities. In Persian white space is used for word boundaries. Another space which is used in Persian Language is ZWNJ space, so-called zerowidth non-joiner, which keeps the word forms close together without joining them. It can cause different challenges in tokenization.

می‌خواهم	using ZWNJ in a right way
می خواهم	using space in a wrong way
میخواهم	using no-space in a wrong way

Table 2: Using white space instead of ZWNJ in word (mixāhæm, I want to)

The freely way of writing multi-word expressions as attached or detached completely distinct word in the multi-word (be hādāfe, in order to):

به‌منظور	using ZWNJ in a right way
به منظور	using space in a wrong way
بمنظور	using no space/omitting one letter in a wrong way

Table 3: the attached and detached form of multiword be mænzure (in order to)

The appearance of the inflectional morphemes as bound to the host or free affixes separated by ZWNJ:

کتابخانه‌های	using ZWNJ in a right way
کتاب خانه ها ی	using space in a wrong way
کتابخانههای	using no space in a wrong way

Table 4: The representation of infelectional morpheme separated by ZWNJ

As it can be concluded from the findings by Mohseni and Minaei-bidgoli (2010), morphosyntactic features of Persian words cause two problems: the number of tags is increased in the corpus (586 tags) and the form of the words is changed. This high number of tags debilitates any taggers to work efficiently. From other side the change of word forms reduces the frequency of words with the same lemma; and the number of words belonging to a specific tag reduces as well. This problem also has a bad effect on statistical taggers. The morphological analyzer by removing the problems helps the tagger to cover a large number of tags in the corpus.

Another issue is the using of the ambiguous letter, Hamze. This letter coming from Arabic, is written normally with a carrier letter such as ئ, ؤ, أ, إ. This letter is not used as a preferably spelling but can be found in Persian Text as well. (Seraji, 2015)

»•«

to smell	بوئیدن	بوییدن
down	پائین	پایین

Table 5: An example of wrongly using of Hæmze in Persian text

As Hosseini Pozveh, Monadjemi, and Ahmadi, (2016) mentioned, there is a concept of `Ezafe Kasreh' or simply Ezafe, which connects two words, mostly nouns, the same structure is mostly like the genitive in English. As Hosseini Pozveh et al. (2016) stated, Ezafe is an unstressed vowel that does not have any writing symbol and it is pronounced 'YE or E'. Sometimes `He' or `Ye' hyponyms, are used to identify the case, but it is grammatically incorrect. Ezafe plays the role of "'s" or "of" in English. Parsing this vowel in a sentence is essential to NER in Persian.

The Unicode Standard has characters for Persian called Extended Arabic-Indic, However some softwares use still the Arabic Unicode characters for Persian letters or a combination of western digits and Persian characters. For instance, it is possible to find the letter ک (*kâf)* in the texts written like ك which is an Arabic letter. Therefore, in analyzing Persian texts, this various encoding should be take into consideration. (Megerdoomian, 2000)

Another characteristic of Arabic script of Persian alphabet is the different shape of them depending on their location in a word; if they come in the first, middle or final part of a word or if they are potentially single alphabet.

Persian Alphabet / الفبای فارسی

Detached	Initial	Medial	Final	Roman	Name	Detached	Initial	Medial	Final	Roman	Name
ا	ا	ـا	ـا	á	alef	ص	صـ	ـصـ	ـص	ṣ	sád
ب	بـ	ـبـ	ـب	b	be	ض	ضـ	ـضـ	ـض	ḍ	zád
پ	پـ	ـپـ	ـپ	p	pe	ط	طـ	ـطـ	ـط	ṭ	tá
ت	تـ	ـتـ	ـت	t	te	ظ	ظـ	ـظـ	ـظ	ẓ	zá
ث	ثـ	ـثـ	ـث	th	se	ع	عـ	ـعـ	ـع	‘	ayn
ج	جـ	ـجـ	ـج	j	jim	غ	غـ	ـغـ	ـغ	gh	ghayn
چ	چـ	ـچـ	ـچ	ch	che	ف	فـ	ـفـ	ـف	f	fe
ح	حـ	ـحـ	ـح	ḥ	ḥe	ق	قـ	ـقـ	ـق	q	qáf
خ	خـ	ـخـ	ـخ	kh	khe	ک	کـ	ـکـ	ـک	k	káf
د	د	ـد	ـد	d	dál	گ	گـ	ـگـ	ـگ	g	gáf
ذ	ذ	ـذ	ـذ	dh	zál	ل	لـ	ـلـ	ـل	l	lám
ر	ر	ـر	ـر	r	re	م	مـ	ـمـ	ـم	m	mím
ز	ز	ـز	ـز	z	ze	ن	نـ	ـنـ	ـن	n	nún
ژ	ژ	ـژ	ـژ	zh	zhe	و	و	ـو	ـو	v/ú	váv
س	سـ	ـسـ	ـس	s	sin	ه	هـ	ـهـ	ـه	h	he
ش	شـ	ـشـ	ـش	sh	shin	ی	یـ	ـیـ	ـی	y/í	ye

Table 6: Different shapes of Persian Alphabet (Trinity School)

Conclusion

This paper presented an overview of part of speech tagging of the Latin and non-Latin scripts languages. In addition, this paper presented some of the features, challenges, issues, tools and corpora, which are being used in the part of speech tagging of Persian languages. The significant challenges of the researches done on Persian languages were generally in the tokenization level and mostly as a result of using the Arabic script and its characteristics. This kind of issues are mostly common between Arabic script languages (Arabic, Persian, Urdu) apart from their complete different morphology. This survey indicates that the challenges and issues of different languages can have a lot in common depending on the script they use and the morphological system they follow. By going through the challenges of different languages, one finds even similarities of the ambiguities between the far different languages.

The collaboration between linguistic and computer science can lead to faster solution for the challenges due to the similarities between the languages. The languages with impressively higher native speakers and little progress in NLP and specifically POS tagging can be compared to the languages with better position in this study area by using those features held in common. Although Part of speech tagging in Persian is a young branch of research, not more than 20 years ago, the successful reported results and researches shows a fast growth of this progress. Data exists in different languages and computational linguistic analysis of different languages would gain a better position if the researchers give more attention to the other languages with high native speakers by using their linguistic experts.

References

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., & Oroumchian, F. (2009). Hamshahri: A standard Persian text collection. *Knowledge-Based Systems*, *22*(5), 382-387.

Amtrup, J. W., Mansouri Rad, H., Megerdoomian, K., & Zajac, R. (2000). Persian-English Machine Translation: An Overview of the Shiraz Project. *Memoranda in Computer and Cognitive Science*.

Assi, S. M. (2005). Word Prediction in a Running Text: A Statistical Language Modeling for the Persian Language, poster presented at the Australian Language Technology Workshop, 2005, Sydney University, Australia.

Assi, S. M. (1997). Farsi Linguistic Database (FLDB*), International Journal of Lexicography, 10*(3), 5-10.

Assi, S. M., & Abdolhoseini, M. H. (2000). Grammatical Tagging of a Persian Corpus. *International Journal of Corpus Linguistics*, *5*(1), 69-81.

Azimizadeh, A., Arab, M. M., Quchani, S. R., (2008). Persian Part of Speech Tagger Based on Hidden Markov Model, 9th International Conference on the Statistical Analysis of Textual Data (JADT), USA.

Beeman, O. W. (2005). *Perisan, Dari and Tajik in central Asia.* The National Council for Eurasian and East European Research.

Bijankhan, M. (2004). The Role of the Corpus in writing a Grammar: An Introduction to a Software. *Iranian Journal of Linguistics*, 19.

Bijankhan, M., Sheykhzadegan, J., Bahrani, M. & Ghayoomi, M. (2011). Lessons from building a Persian written corpus: Peykare. *Language resources and evaluation, 45*(2):143–164.

Brants, T. (2000). TNT: A statistical part-of-speech tagger, In the Proceedings of 6th conference on applied natural language processing (ANLP), USA.

Brill, E. (1995). Transformation-based Error driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging, *Journal of Computational Linguistic*, 21(4), 543-565.

Curtis J. E. and Tallis, N. (2005). *Forgotten Empire, The World of Ancient Persia,*.University of California Press, Berkeley and Los Angeles, California.

Fadaei, H. & Shamsfard, M. (2008). Persian POS tagging using probabilistic morphological analysis. *International Journal of Computer Applications in Technology, 38*(4), 264-273.

Forsati, R. & Shamsfard, M. (2012). Cooperation of evolutionary and statistical POS-tagging. In The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), pages 446-451.

Hashemi, H. B., Shakery, A. & Faili, H. (2010). *Creating a Persian-English Comparable Corpus*, in proceedings of Conference on Multilingual and Multimodal Information Access Evaluation (CLEF), Padua, Italy, pp. 27-39.

Hosseini Pozveh, Z., Monadjemi, A., Ahmadi, A. (2016). Persian Texts Part of Speech Tagging Using Artificial Neural Networks, *Journal of Computing and Security, 3*(4). 233-241.

Jurafsky D., & Martin, J. H. (2009). *Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition,* Prentice-Hall, Upper Saddle River, New Jersey.

Kardan, A. A. & Imani, M. B. (2014). Improving Persian POS tagging using the maximum entropy model. In 2014 Iranian Conference on Intelligent Systems (ICIS), 1-5.

Karlsson, F., Voutilainen, A., Heikkilä, J. & Anttila, A. (1995*). Constraint Grammar: A Language-Independent Framework for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin /New York.

Khanam, M. H., Madhumurthy, K. V., Khudhus, M. A. (2013). Part-Of-Speech Tagging for Urdu in Scarce Resource: Mix Maximum Entropy Modelling System, *International Journal of Advanced Research in Computer and Communication Engineering,* 2(9).

Khoja, S. (2001). *Arabic Part of Speech Tagger,* Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001), Carnegie Mellon University, Pittsburgh, Pennsylvania.

Megerdoomian, K. (2004). Developing a Part of Speech Tagger. *In Proceedings of First Workshop on Persian Language and Computers*. Iran.

Mohseni, M., & Minaei-Bidgoli, B. (2010). A Persian Part-of-Speech Tagger Based on Morphological Analysis.*The International Conference on Language Resources and Evaluation*, Valletta, Malta.

Mohseni, M., Motalebi, H., Minaei-bidgoli, B., Shokrollahi-far, M. (2008). A farsi part-of-speech tagger based on markov, In the proceedings of ACM symposium on Applied computing, Brazil.

Mohtarami, M., Oroumchian, F. & Rahgizar, M. (2008). Using Heuristic Rules to Improve Persian Part of Speech Tagging Accuracy*, International Conference on information and Knowledge Engineering*, Universal Conference Management Systems and Support, California, USA.

Ng, H. T. & Low, J. K. (2004). *Chinese Part-of Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based?,* Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 277-284.

Nourian, A., Rasooli, M. S., Imany, M. & Faili, H. (2015). On the importance of ezafe construction in Persian parsing. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), volume 2, 877-882.

Okhovvat, and Minaei Bidgoli, B. (2011). A Hidden Markov Model for Persian Part-of-Speech Tagging, *Procedia Computer Science 3*, 977–981.

Pakzad, A. & Minaei Bidgoli, B. (2016). An improved joint model: POS tagging and dependency parsing. *Journal of AI and Data Mining, 4*(1), 1-8.

Raja, F., Amiri, H., Tasharofi, S., Sarmadi, M., Hojat, H. (2007). Evaluation of Part of Speech Tagging on Persian Text. *Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages Linguistic Institute,* Stanford, California, USA, pp. 21-22.

Rehman, Z. Anwar, W., Bajwa, U. I. Xuan, W., & Chaoing, Z. (2013). *Morpheme Matching Based Text Tokenization for a Scarce Resourced Language,* PLoS ONE 8(8): e68178. Retrived from https://doi.org/10.1371/journal.pone.0068178.

Schuetze, H. (1995). Distributional Part-of-Speech Tagging From Texts to Tags: Issues in Multilingual Language Analysis, In the Proceedings of the ACL SIDGAT Workshop, available at: http://xxx.lanl.gov/find/cmp-lg.

Seraji, M., Megyesi, B., & Nivre, J. (2012). Dependency parsers for Persian. In Proceedings of the 10th Workshop on Asian Language Resources, 35-44.

Seraji, M. (2015). *Morphosyntactic Corpora and Tools for Persian*, Uppsala University, Sweden.

Sheykh Esmaili, K., Abolhassani, H., Neshati, M., Behrangi, E., Rostami, A., & Mohammadi, M. (2007). Mahak: A Test Collection for Evaluation of Farsi Information Retrieval Systems, IEEE/ACS International Conference on Computer Systems and Applications.

Supreme Council of Information and Communication Technology, Mizan English Persian Parallel Corpus, (2013). Available: http://dadegan.ir/catalog/mizan [2014-01-01].

Taghiyareh F., Darrudi E., Oroumchian F., Angoshtari N. (2003) Compression of Persian Text for Web-Based Applications, Without Explicit Decompression, WSEAS Transactions on Computers, 4 (2), 961-966.

Vikram, S. (2013). Morphology: Indian Languages and European Languages. *International Journal of Scientific and Research Publications, 3*(6), 1-5.

Xia, F. (2000). *The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank*, University of Pennsylvania.

HOW TO CITE THIS ARTICLE

Moghadam, M., & Jafarpour, N. (2021). A Survey of Part of Speech Tagging of Latin and non-Latin Script Languages: A more vivid view on Persian. *Language Art*, 6(1): 75-90, Shiraz, Iran.

DOI: 10.22046/LA.2021.05

URL: //www.languageart.ir/index.php/LA/article/view/180

Language Art, 6(1): pp.75-90, 2021, Shiraz, Iran
DOI: 10.22046/LA.2021.05 DOR: 98.1000/2476-6526.1399.6.75.18.1.65.110
فصلنامه هنر زبان، دوره ۶، شماره ۱، سال ۲۰۲۱، از صفحه ۷۵ تا ۹۰

مروری بر برچسب‌گذاری واژگانی زبان‌هایی با صورت نوشتاری لاتین و غیرلاتین: نگاهی مبسوط بر زبان فارسی

دکتر میثم مقدم[1]©

استادیار، دانشکده علوم، دانشگاه فسا، ایران

نیلوفر جعفرپور[2]

کارشناسی ارشد زبان‌شناسی و ادبیات رایانشی، دانشکده زبان‌شناسی و مطالعات ادبی
دانشگاه تکنیکی دارمشتات، آلمان.

(تاریخ دریافت: ۱۴ خرداد ۱۳۹۹؛ تاریخ پذیرش: ۸ آذر ۱۳۹۹؛ تاریخ انتشار: ۱۰ اسفند ۱۳۹۹)

مقاله حاضر، به بررسی جامع موضوع برچسب‌گذاری واژگانی صورت نوشتاری زبان‌های لاتین و غیرلاتین به ویژه زبان فارسی می‌پردازد. در این نوشتار میزان پیشرفت برچسب‌گذاری واژگانی در بیست و سه زبان گفتاری دنیا، که دارای بیشترین متکلم می‌باشند، مورد بررسی قرار می‌گیرد. برخی از این زبان‌ها مثل زبان‌های عربی، اردو و فارسی از سیستم نوشتاری از راست به چپ پیروی می‌کنند، و در نوع خود با مشکلات و چالش‌هایی در زمینه برچسب‌گذاری واژگانی روبرو هستند. این چالش‌ها می‌تواند منحصر به یک زبان خاص باشد و یا در بین زبان‌های گوناگون مشترک باشند، که به برخی از آن‌ها اشاره خواهیم کرد. در این مقاله، با مروری نقادانه بر مطالعات اخیر در حیطه برچسب‌گذاری واژگانی، چالش‌های پیش روی زبان فارسی مد نظر قرار گرفته شده است. با مرور تحقیقات پیشین و مطالعه ویژگی‌ها، مسائل، چالش‌ها و ابزارهای برچسب‌گذاری واژگانی، این نتیجه حاصل می‌شود که، چالش‌های برچسب‌گذاری واژگانی در زبان فارسی بیشتر در سطح توکن‌سازی و مربوط به شرایط رسم الخط عربی است.

واژه‌های کلیدی: برچسب‌گذاری واژگانی، زبان‌های نوشتاری لاتین، زبان‌های نوشتاری غیر لاتین، زبان فارسی، سیستم RTL.

©(نویسنده مسؤول)

[1] E-mail: moghaddam.m@fasau.ac.ir

[2] E-mail: niloofar.jafarpour@yahoo.com