The Problem of Pos Tagging and Stemming for Agglutinative Languages

(Turkish, Uyghur, Uzbek Languages)

Elov Botir Boltayevich
Dept. of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named Alisher Navo'i
Tashkent, Uzbekistan
elov@navoiy-uni.uz

Eşref Adalı
Computer Engineering and Informatics Faculty
Istanbul Technical University
Istanbul – Türkiye
adali@itu.edu.tr

Khamroeva Shahlo Mirdjonovna
Dept.of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named Alisher Navo'i
Tashkent, Uzbekistan
shaxlo.xamrayeva@navoiy-uni.uz

Abdullayeva Oqila Xolmo'Mınovna
Dept.of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named Alisher Navo'i
Tashkent, Uzbekistan
abdullayevaoqila@gmail.com

Xusainova Zilola Yuldashevna
Dept. of Computational Linguistics and Digital Technologies
Tashkent State University of Uzbek Language and Literature named Alisher Navo'i
Tashkent, Uzbekistan
xusainovazilola@navoiy-uni.uz

Xudayberganov Nizomaddin Uktamboy O'g'li
Dept.of Computational Linguistics and Digital Technologies
Tashkent State Univ.of Uzbek Language and Literature named Alisher Navo'i
Tashkent, Uzbekistan
nizomaddin@navoiy-uni.uz

***Abstract*—The number of possible word forms in agglutinative languages is theoretically unlimited. This, in turn, creates the problem of POS tagging (part-of-speech) of out-of-vocabulary (OOV) words in agglutinative languages. In agglutinative languages, words are formed by adding suffixes to the stem. Due to the occurrence of phonetic harmony and disharmony while adding suffixes to the stem, it is necessary to analyze both phonetic and morphological changes. When solving many NLP tasks, it is necessary to reduce word forms to the stem (stemming). Removing all inflectional affixes from a word and lemmatizing the rest of the word is considered one of the important tasks of natural language processing (NLP), and this process is called stemming. The stemming process is important in information retrieval (IR) systems.**

Keywords—part-of-speech, POS tagging, stemming, information retrieval, IR, stemming algorithms

I. INTRODUCTION

Increasing the speed of returning a result that matches the user's query is one of the most important issue in information retrieval systems. The easiest and most convenient way to do this is through the stemming process. In NLP, the method that determines the general form (stem) of various morphological variants of a word is called the stemming algorithm [1]. To identify the stem in information retrieval systems, it is necessary to remove its suffixes and prefixes [2]. POS tagging is the task of determining (tagging) which type of words (noun, verb, adjective, number, adverb, or pronoun) belongs to each word in a given sentence. POS labeling is one of the main tasks of natural language processing (NLP) and an important pipeline step (Figure 1).

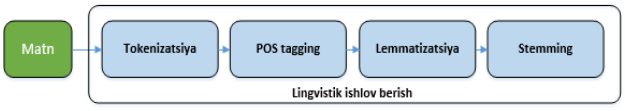

Figure 1. Stages of initial text processing

II. METHODS

POS tagging is an essential step for NLP applications such as machine translation, text summarization, question-answering, and sentiment analysis, For example, the POS tagging is used to translate the word "olma" (an apple) into other languages. "Olma" (an apple) is an object if it belongs to the noun group, and if "ol-ma" (don't take)" it indicates an action and belongs to the verb group. POS tagging can be done with or without a dictionary. Most of the scientific studies on POS tagging are word-based and do not perform morphological segmentation of words [3]. Some agglutinative languages use word stems to implement the POS tagging process [4]. Uzbek, Turkish and Uyghur words and their stems may belong to different POS tags.

TABLE 1. LEMMA, STEM AND POS TAG OF WORD FORM IN AGGLUTINATIVE LANGUAGES

Word	lemma	POS	Stem	POS	stem	POS
Muzladi	muzlamoq	VB	muz	N	muz	N
Issiqroq	issiq	JJ	isi	VB	isi	VB
Soddalashtiriladi	soddalashtirmoq	VB	sodda	JJ	sodda	JJ
Ixtiyoriy	ixtiyoriy	JJ	ixtiyor	N	ixtiyor	N
qo'llaniladigan	qo'llamoq	VB	qo'l	N	qo'l	N
yo'lakda	yo'lak	N	yo'l	N	yo'l	N
qishlog'im	qishloq	N	qishlog'	?	qishloq	N
Yetkili	yetkili	ADJ	yetkili	ADJ	yetki	N
Kurullarimizla	kurul	N	kurul	N	kurul	N
Teşkilatlarimizla	teşkilat	N	teşkilat	N	teşkil	N
seçimlere	seçim	N	seçim	N	seç	VB
futbolcularin	futbolcu	N	futbolcu	N	futbol	N

Kullandi	kullanmak	F	kulla	F	kulla	F
bilgi	bilgi	N	bilgi	N	bil	F
tarazichi	tarazichi	N	tarazichi	N	tarazi	N
yashaptu	yashamaq	VB	yasha	VB	yash	N
yegizligi	yegizlik	N	yegizlig	N	yegiz	VB
og'urluqqa	og'urluq	N	og‘urluq	N	og‘ur	N
chγshkənligini	chγshkənliq	N	chγshkən lig	N	chγsh	?

The process of stemming in Turkish and Uyghur is described as follows:

Stemming (Turkish and Uyghur) is the process of reducing a word to its core by removing inflectional suffixes. Table 2 below lists words in Uzbek, Turkish, and Uyghur languages, their stems, and examples of word-forming and form-forming suffixes added to their stems which is shown in Table-II.

TABLE II. STEMS AND SUFFIXES OF THE WORD FORM IN AGGLUTINATIVE LANGUAGES

Til	So‘zshakl	Stem	So‘z yasovchi qo‘shimcha	Shakl yasovchi qo‘shimcha
UZ	ko‘zlagan = ko‘z + la + gan	ko‘z	la	gan
	tinchimiz = tin + ch + imiz	tin	ch	imiz
	bilimdon = bil + im + don	bil	im	don
	birlik = bir + lik	bir	lik	bir
	moyladim = moy + la + di	moy	la	di +m
TR	oyuncularin = oyun+cu+lar+in	oyun	cu	lar+in
	futbolcularin = futbol+cu+lar+in	futbol	cu	lar+in
	karşilaşmalar = karşi+laş+ma+lar	karşi	laş	ma+lar
	değerlendirilip = değer+len+dir+il+ip	değer	len	dir+il+ip
	açikladi = açik+la+di	açik	la	di
UY	tarazichi = tarazi+chi	tarazi	chi	-
	yashaptu = yasha+p+tu	yash	a	p+tu
	yegizligi = yegiz+lig+i	yegiz	lig	i
	og'urluqqa = og‘ur+luq+qa	og‘ur	luq	qa
	chγshkənligini = chγshkən+lig+i+ni	chγshkən	lig	i+ni

However, the stemming process for the Uzbek language is described as follows:

Stemming (Uzbek) is the task of reducing the word to its core by removing the derivational and inflectional suffixes added to it. In Uzbek, Turkish and Uyghur languages, sentences consist of separate words. Morphologically, words in these three languages are formed by adding some suffixes to the root. In this process, phonetic changes (phonetic harmony) may occur in the word, and this is directly reflected in the text. The root itself can also be a word that expresses the specific meaning of the word. Although affixes play an important role in the sentence, they do not have an independent meaning.

Affixes are divided into derivational suffixes and inflectional suffixes [5]. In Turkish and Uyghur, word-forming suffixes can form new stems (Fig. 2). Form-forming suffixes change only the grammatical function of the word. A semantic change can occur in a word by adding word-forming suffixes to the stem. Form-forming suffixes cause syntactic changes in the word. Word-forming suffixes are added to the root first, and then form-forming suffixes. However, it is also possible to add form-forming suffixes directly to the stem.

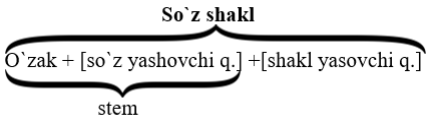

Fig. 2. The general morphological structure of the word in the Turkish and Uyghur languages

In Turkish and Uyghur, roots together with word-forming suffixes turn into stems. In agglutinative languages, form-forming suffixes usually follow word-forming suffixes. However, in some cases, form-forming suffixes such as -gil, -siz can come first.

In the Uzbek language, a lexical form is preceded by a word-former, and as an example, we can cite the words o‘chirg‘ich, muzlatkich.

o‘chirg‘ich = o‘ch (root) + ir (lexical form-former) + g‘ich (word former)

muzlatkich = muz +la (word-former) +t (lexical form-former) +kich (word-former)

In the Turkish language, after the root, word- forming suffixes + lexical form- forming suffixes + word- forming suffixes form is found:

baş+la+n+gıç;

There is also a root + syntactic form- forming suffixes + word- forming suffixes form:

aşağıdaki (sorular), aşağıdakiler, sınıftaki (öğrenciler), sınıftakiler, raftaki (eşyalar), yuvadaki [6].

In the Uyghur language, there are also words that do not correspond to the order of stem + word- forming suffixes + form- forming suffixes, i.e., stem + form- forming suffixes + word- forming suffixes:

oqu+t-quchi; qolla+n-ma[7].

The number of suffixes that can be added to a word and their numerous combinations make the of root identifying process in agglutinative languages a complex problem. Because in most agglutinative languages, combinations of suffixes form complex word forms.

As can be seen from Table 2 above, Uzbek, Turkish and Uyghur languages look at stem and lemma differently. In the Uzbek language, lemma is in the form of a root or artificial word: book, book reader, knowledge, scholar. So, in Uzbek language, lemma is equal to lexeme in the dictionary. In the Uzbek language, cognate (base) words are counted as lemmas separately. In order to perform stemming in the Uzbek language, all suffixes up to the root of the word form are cut off. In a word form Maktab+dosh+lar+imiz there is a word-forming and a form-forming suffix. In the process of stemming in Uzbek, all these suffixes are cut off:

Word form: *maktab*+{dosh}+(lar)+(imiz)
Lemma: *maktabdosh*
Stem: *maktab*
Root: *maktab*

In the process of stemming in Turkish, only syntactic and lexical form-forming suffixes in the word form are cut, but the word-forming ones are left. For example:

Word form: seçim+(ler)+(e)
Stem: seçim

It can be seen that in Turkish, a word-forming suffix remains in the stem, the difference between root and stem is the presence of a word-forming suffix.

Word form: seçim+(ler)+(e)
Lemma: seçim
Stem: seçim
Root: seç

In the process of stemming in the Uyghur language, the syntactic and lexical form-forming suffixes in the word form are cut, but the word-forming suffixes are left.

oqutquchi
Word form: oqut + (qu) + (chi)
Lemma: oqut
Stem: oqut
Root: o

III. RESULTS AND DISCUSSION

POS tagging of corpus texts is widely used as a clustering problem in NLP. Brown presented a class-based n-gram model based on a complex hierarchical clustering algorithm for learning syntactic classes of words [8]. In the study, the context information is entered in the form of n-grams, and in the initial state, each word belongs to one group. Then, each pair of clusters that gives the average minimum loss is aggregated until all clusters are merged under one cluster. In the next step, a binary tree representing the hierarchy between syntactic categories is formed.

Banko and Moore provide a contextual HMM tagger for each word based not only on the current word "tag" but also on three neighboring tags, including previous and subsequent word "tags" (valence) [9]. Compared to the basic HMM, this model included more contextual information and showed efficient results.

Johnson compared the various parameters used in HMM-based POS tagging. For this purpose, he used Expectation Maximization (EM), Variational Bayes and Gibbs sampling [10]. The study showed the low efficiency of EM algorithm compared to Gibbs sampling and variational Bayes estimator.

Researches on stemming. Modern stemming algorithms are generally divided into three classes: rule-based, statistical, and hybrid algorithms (Fig. 7). Rule-based stemmers aim to identify stems using non-automatic rules. Popular rule-based stemmers include Lovins [11], Porter [12] and Krovets [13]. Rule-based stemming algorithms are usually controlled.

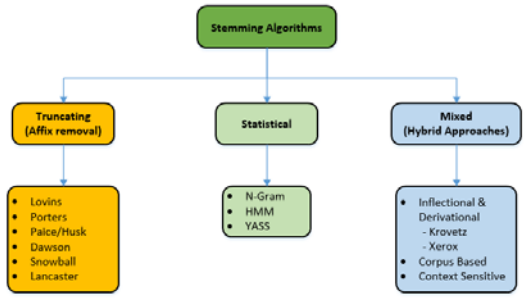

Fig. 3. Classification of stemming algorithms

Statistical stemming algorithms use statistical methods to learn stems. Xu and Croft [14] presented a method that uses a random statistical term to overcome the shortcomings of Porter's stemming. Based on the data of random statistics, they applied a graph partitioning algorithm to reduce the number of classes generated by Porter's stemmer [15].

Hybrid stemming algorithms combine rule-based and statistical methods into a single system. Some hybrid stemming algorithms have been developed by Shrivastava [16], Goweder [17] and Adam [18].

A method for implementing Turkish stemming was introduced by Köksal [8]. This method is based on considering the first 5-6 letters as the root. Kut et al. developed a method called L-M (Longest Match) in their research. Using a dictionary containing word stems and their possible forms, the method compares the stem word with the words in the dictionary from left to right. The longest matching word is the stem.

Solak and Can [1] used a dictionary of roots to identify stems. Each stem is recorded as having 64 features corresponding to the left-to-right stemming methods. Letter units are mapped to the root lexicon in left-to-right order, and if a matching stem is found, the system determines possible stems based on additional rules. This research, called the AF algorithm, is basically a variant of the morphological analysis method developed by Oflazer.

FindStem is a stemming method developed by Sever and Bitirim [3], which mainly includes three steps: stem detection, stem morphological analysis, and detection. The method uses a dictionary containing morphological and POS features of words, syntactic rules. Sever and Bitirim claim that the FindStem algorithm works better and more efficiently than the AF and L-M algorithms.

Other analytical methods for determining the stem of Turkish words include the "zemberek" algorithm developed by Akin [19] and the "snowball" algorithm developed by Childen [3]. Also, Dincher [20] proposed a method for solving the boundary between the root and the suffixes using n-gram statistics. As a result of the application of this research, the efficiency was 95.8%.

Aishan Wumaier and other researchers developed a new Uyghur noun stemming method in 2009 [21]. The Uyghur noun stemming method was implemented in 2 stages:

- Uyghur language using FSM additions;
- Using the CRF method to eliminate ambiguities caused by Uighur FSM suffixes.

In the first stage, the process of Uyghur noun stemming was developed using FSM noun suffixes. The stemming process was performed on 55,625 input words, and as a result, 6,239 incorrect over-stemming words were identified. In the second stage, 55125 word corpus was built by determining the inaccuracies that occurred during the stemming process using the CRF method. The corpus consists of 17317 words with indefinite adverbs, 6239 words without correct adverbs and 11078 words with correct adverbs. The result of the algorithm shows that the recall rate was 88.78% when FSM additives were used, and the recall rate was 94.04% when FSM and CRF were used. In conclusion, using the CRF method improves the recovery rate by 5.26%.

In 2012, Azragul, Qixiangjwei, and Yusupulla developed a Uyghur language stemmer [22]. They used a dictionary-based method. During the operation of the algorithm, the entered word is searched from the stem dictionary. In this case, a word is separated using a dictionary of suffixes, and a candidate word separated by removing the suffixes is searched in the dictionary.

Studies have shown that previous studies used an incomplete vocabulary (open vocabulary) and the inaccuracies resulting from stemming were subsequently resolved by other methods.

IV. STUDY OF THE PROBLEM

4.1. Problems In The Stemming Process

The following issues may occur during stemming:

1) the stem and suffix are homonymous with one stem;
2) the occurrence of a sound change in the word;
3) stemming neologism and NERs.

The stem and suffix are homonymous with one stem. Today, various stemming methods have been developed for natural language words. Modern stemming algorithms are being developed without using any syntactic information.

Also, traditional stemming methods (algorithms) are based on suffixes and some morphological rules, and as a result of the stemming process, ambiguity in the stem may occur. Determining a polysemous stem is a more complex process, and sentence-level semantic information is ignored in the stemming process. Sometimes the POS tag of a word may not be the same as the POS tag of its root.

In the Uzbek language, there is also a phenomenon of homonymy between word-forming and form-forming suffixes. This creates a problematic situation in the stemming process. Table III below provides a list of homonyms:

TABLE III. HOMONYMY BETWEEN WORD-FORMING AND FORM-FORMING SUFFIXES

Shakl yasovchi qo‘shimcha	**So‘z yasovchi qo‘shimcha**
-ay (lug‘aviy shakl yas.) boray	kuchay (fe’l)
-gi (lug‘aviy shakl yas.) borgim	supurgi (ot), yozgi (sifat)
-da (sintaktik shakl yas.) uyda	undamoq (fe’l)
-i (sintaktik shakl yas.) do‘sti	jannati (sifat), boyi (fe’l)
-in (lug‘aviy shakl yas.) ko‘rin	ekin (ot), sog‘in (sifat)
-im (sintaktik shakl yas.) uyim	bilim (ot), ayrim (sifat)
-ir (lug‘aviy shakl yas.) o‘chir	gapir (fe’l)
-iq (lug‘aviy shakl yas.) siniqmoq	yo‘liq (fe’l), ochiq (sifat), chiziq (ot)
-y (lug‘aviy shakl yas.) o‘qiy	qoray (sifat)
-k (sintaktik shakl yas.) bordik	to‘shak (ot), chirik (sifat)
-ka (lug‘aviy shakl yas.) surka	iska (fe’l)
-kin ((lug‘aviy shakl yas.) to‘kkin	epkin (ot), keskin (sifat)
-la (lug‘aviy shakl yas.) quvla	so‘zla (fe’l)
-lab (lug‘aviy shakl yas.) yuzlab	haftalab (ravish)
-m (sintaktik shakl yas.) otam, ko‘rdim	to‘plam (ot)
-ma (lug‘aviy shakl yas.) gapirma	qatlama (ot), bo‘g‘ma (sifat)
-moq (lug‘aviy shakl yas.) ichmoq	quymoq (ot)
-sa (lug‘aviy shakl yas.) kelsa	suvsa (fe’l)
-siz (sintaktik shakl yas.) yozasiz	yuzsiz (sifat), to‘xtovsiz (ravish)

TABLE IV. STEMS AND SUFFIXES OF WORD FORM IN AGGLUTINATIVE LANGUAGES

Til	So‘zshakl	1-ma‘nosi	2-ma‘nosi
Turk	gelecek	keladi (will come)	kelajak (future)
Uyg‘ur	alma	ol+ma (don’t take)	olma (apple)
O‘zbek	quymoq	quy+moq (pour)	quymoq (panke)

The ambiguity of the stem in Uzbek language sentences can be seen in the following Figure 3:

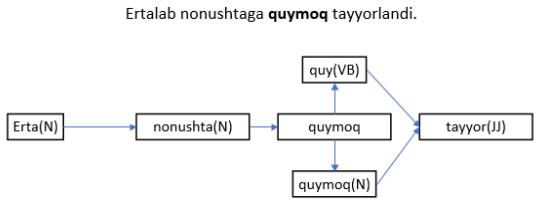

Fig. 4. Stem ambiguity in Uzbek language sentences

In Turkish

1) meaning: *Kış yine gelecek*. (Winter will come again)
2) meaning: *Gelecek hakkında ne düşünüyorsunuz*? (What do you think about the future?)

In Uyghur:

1) meaning: *Qalamni qolunga olma*. (Don't take your pen.)
2) meaning: *U bazardin olma setiwaldi*. (He sold apple on the market.)

In Uzbek: (quymoq)

1) meaning: *Zarifa mehmonlarga choy quymoqchi bo‘ldi*. (Zarifa wanted to pour tea for the guests.)
2) meaning: *Ertalab nonushtaga quymoq tayyorlandi*. (In the morning, pudding was prepared for breakfast.)

Various problems can occur during the POS tagging process. One of them is the ambiguity in POS tagging. Words can belong to different word groups depending on their syntactic role in a sentence. The correct POS tag of a word helps to find its stem. For example, in Turkish

1) *Aydınlık gelecek günler bizi bekliyor. (Brighter days await us in the future).*
2) *Ahmet birazdan gelecek. (Ahmat is coming soon);*

gelecek in the first sentence is an adjective, and the root is *gelecek* (future). In the second sentence, gelecek is a verb, and the stem is gel-(mek) (to come). From the above considerations, it can be noted that POS tagging process plays an important role in stemming.

We can observe a similar situation in the Uyghur language. For example, the word *olma* is stemmed in the form of apple in the meaning of *apple fruit*, and ol-ma as a verb is stemmed in the form of olma. In stemming, the difference in word forms from POS tagging can also be observed in the word kelgüsi.

1) *Kelgüsi ishimni planladim. (I planned my future work)*
2) *Bala ete kelgüsi. (Child is coming soon)*

In the first sentence *kelgüsi* is an adjective, and the stem is *kelgüsi* (the future). In the second sentence, kelgüsi is a verb in the future tense, and the stem is in the form kel-(mek) (to come).

In the Uzbek language, the stem and suffix can be homonymous with one stem, and the complications in POS tagging and stemming can be observed in many examples. For example, *tortma, olma, yozma, o‘sma* and etc are word forms. These words are in the form of stem *tortma* - tort-(moq), *olma*-ol-(moq), *yozma*-yoz-(moq), *o‘sma*-o‘s-(moq), and POS tagging is defined as a noun and a verb. For example:

1) *Sen bozordan kitob olma (Don't buy books from the market)*
2) *Akbar kecha olma yedi. (Yesterday Akbar ate an apple)*

Here, in the first sentence, olma is a verb with a negative meaning, and the stem is in the form of ol-(moq), in the second sentence, olma is a noun, and the stem is also olma. From the above considerations, it can be seen that in all three Turkish languages, there is a case where the root and the suffix are homonymous with one root, and in this case, it can be noted that the process of POS tagging plays an important role in stemming.

The Occurrence of a Sound Change in The Word

Phonetic changes (insertion, deletion, phonetic harmony, and assimilation) may occur in some cases as a result of adding form-forming suffixes to the last letters of the stem. In agglutinative languages, three types of phonetic changes can be made in a word, such as sound increase, decrease and exchange (weaking, assimilation). (Table V).

TABLE V. DEFICIENCIES IN THE STEMMING PROCESS IN AGGLUTINATIVE LANGUAGES

O‘zbek		Turk		Uyg‘ur	
to‘g‘ri	***xato***	***to‘g‘ri***	***xato***	***to‘g‘ri***	***xato***
lavozim+ida ish+lagan	boshlig‘i san+aydi	yara+landi ğini belirt+ti	ögre-nlere jandar+ maliğin a	yoğ+an eshek+ medeği	binay+im oghl+um
hafta+larida	tarog+`ini	ara+sinda	rastla+d iği	chaplish +ivalid u	yot+im
bo‘lim+i hokim+ining ish+lagan	me+ning obro‘y+im iz achch+iq	koşul+lard a gösteri+ci nin belir+li	iznin+e geti+rili yor	bash+la pti dep+ti ini+sini n	yurag+im shahr+im

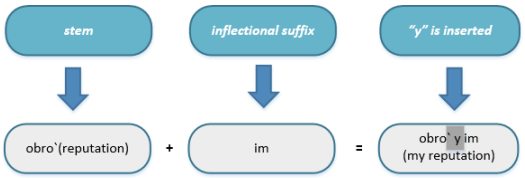

Fig. 5. Increase in sound in the word

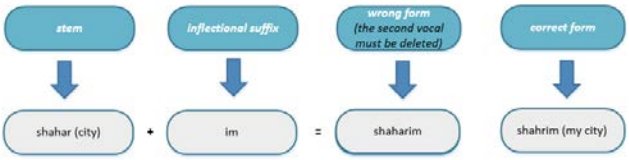

Fig. 6. Decreasing of sound in a word

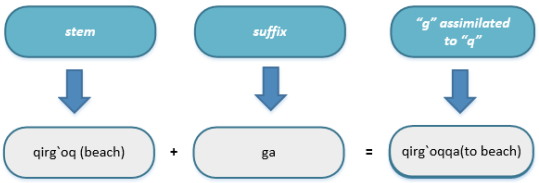

Fig. 7. Sound exchange in a word

In order to solve the problem of sound change in stem detection, the boundaries of the stem and affix are determined in the first step, and the lemmatization is performed in the second step. As a result of lemmatization, the wrong stems are changed to the root in the dictionary.

4.2. Stemming Neologisms and Ners. Problems of Stemming Ners

The suffix -lik is derived mainly from nouns, adjectives, and adverbs. Words made from lexemes related to the noun family have different meanings depending on the nature of the object represented by the constituent base:

1) when it is made from words denoting a person (*otalik, onalik, tog‘alik, o‘g‘illik, farzandlik, erlik, xotinlik*;) Nouns made from words denoting a certain period of life (*bolalik, yigitlik, qizlik, o‘smirlik, kelinlik, kuyovlik*)nouns made from words denoting the owner of a profession or title (*mudirlik, o‘qituvchilik, qassoblik, chorvadorlik, tabiblik, suvchilik, sartaroshlik, savdogarlik, rassomlik, shofyorlik, aktyorlik*);
2) a noun denoting the object occupied by the thing understood from the base (*botqoqlik, qumlik, muzlik*);
3) a noun denoting the part of the earth's surface understood from the constituent base (*jarlik, do‘nglik, qiyalik, pastlik, ichkarilik, yalanglik*).

By combining adjectives and adverbs, a noun is formed: (*qizillik, semizlik, xursandlik, aniqlik*) In such cases, the word-forming suffixes in their content are cut, and the remaining part is considered as a stem. But when the suffix -lik is added to the proper nouns denoting the name of a place, they become a common noun and is written with a lowercase letter: (samarqandlik, buxorolik, amerikalik, o‘zbekistonlik, turkiyalik, arabistonlik). In this case, the suffix - *-lik* cut off, the remaining part is understood as a stem, converted to a capital letter and recognized as NER.

samarqandlik = Samarqand~~lik~~
amerikalik = Amerika~~lik~~
kanadalik = Kanada~~lik~~

When there is a problem of finding the stem of NERs, the form-forming suffixes are cut off, the suffix of the word-forming form or part of the word is left, this part is considered NER: the stem of the word form O‘zbekistondan is O‘zbekiston. There are suffixes that have functions as word-forming suffix and form-forming one (Table VI).

TABLE VI. WORD-FORMING AND FORM-FORMING HOMONYMOUS SUFFIXES

the form-forming and word-forming suffixes		
-ay	-k	-chak
-gi	-ka	-chiq
-da	-kin	-choq
-i	-la	-qa
-in	-lab	-qin
-im	-m	-sa
-ir	-ma	-siz
-iq	-moq	-xon
-y	-cha	

When such suffixes appear in the composition of words written with a capital letter, if there are form forming suffixes and word-forming suffixes in its composition, the word remains in the composition of the form and is considered a stem in this form. For example, the word Jon Kennedi (John Kennedy) contains the letter -i. Since the program does not know the stem, that is, the word does not exist in the dictionary of the Uzbek language, it cannot distinguish the stem, as a result, it can cut the suffix -i and take the word Kenned as the stem. In order to avoid such a situation, any unit that is cognate with a suffix that creates homonymy between the form-maker and word-formers is left in the word-form.

4.3. Problems of Stemming Neologisms

Neologism expressing new things and concepts that appeared with the development of society, the needs of life. The novelty of neologisms is noticeable only at the time of their initial appearance: over time, they lose the "novelty" feature and usually become active words. There are types of neologism such as formal neologism, semantic neologism, functional neologism, social neologism, technological neologism, stylistic neologism. There are different ways of neologisms, they are created by creating a new word based on the existing lexical structure of the language and grammatical rules, as well as by using one of the dictionary meanings of the existing word in a new sense and by adopting a word from another language. Neologisms include suffixes such as - ism (neologism), -ik (daltonik), -la (gugllash).

Since neologisms are not in the dictionary, problems arise in their stemming. Among them are the additions in their composition, the problems of a part of the word resembling a suffix. In this case, existing suffixes in the database of form-forming additions will be cut. The remaining part corresponds to the stem. The stem corresponding to neologisms and NERs

in the Uzbek language corresponds to the definition of the stem in the Turkish and Uyghur languages given in Figure 2 above.

V. CONCLUSION

The implementation of POS tagging and stemming through a dictionary is a challenge for many natural language processing tasks. Using a language corpus for POS tagging and stemming solves problems with vocabulary. Various experiments on language corpora show that combining stem information with a syntactic task improves the POS tagging result for a morphologically rich language, which improves the solving efficiency of the NLP task. In the article, several different joint models are presented, which assume different dependencies. Overall experimental results show that the Bayesian HMM model using neural word embeddings outperforms other models for the POS tagging task. Also, when using the semantic similarity between the stem and the words to determine the inflectional morphology, the inflectional suffixes do not change the meaning of the word. For this purpose, the method of neural word embeddings obtained from word2vec should be used. The results show that using semantic information significantly improves stemming and POS tagging.

REFERENCES

[1] C. D. Paice, An evaluation method for stemming algorithms. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1994. https://doi.org/10.1007/978-1-4471-2099-5_5

[2] M. Anjali and G. Jivani, A Comparative Study of Stemming Algorithms. Int. J. Comp. Tech. *Appl.*, *2*(6). 2011.

[3] S. Goldwater , T. L. Griffiths, A fully Bayesian approach to unsupervised part-of-speech tagging. ACL 2007 - Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 2007.

[4] B. Taner Dinçer and B. Karaoğlan, Stemming in agglutinative languages: A probabilistic stemmer for Turkish. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2869, 2003 https://doi.org/10.1007/978-3-540-39737-3_31

[5] А. Ҳожиев, Ўзбек тилида сўз ясалиши. − Ташкент, 2005.

[6] https://www.turkedebiyati.org/yapim-ekleri/

[7] https://tilachar.ru/ru/grammar/24-grammar

[8] P.Brown, V.DellaPietra, de Souza, P., J.Lai, R.Mercer, Class-based n-gram models of natural language. Computational Linguistics, 18., 1992.

[9] M.Banko, R.C.Moore, Part of speech tagging in context. COLING 2004 - Proceedings of the 20th International Conference on Computational Linguistics. https://doi.org/10.3115/1220355.1220435, 2004.

[10] A. Haghighi and D. Klein, Prototype-driven learning for sequence models. HLT-NAACL 2006 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings of the Main Conference, 2006 https://doi.org/10.3115/1220835.1220876.

[11] J. B. Lovins, Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11(June), 1968.

[12] M.F. Porter, “Snowball: A language for stemming algorithms”, 2001.

[13] R. Krovetz, Viewing morphology as an inference process. Artificial Intelligence, 118 (1–2), 2000. https://doi.org/10.1016/S0004-3702(99)00101-0.

[14] J. Xu, W. B. Croft, Corpus-based stemming using cooccurrence of word variants. ACM Transactionson Information Systems, 16(1)., 1998. https://doi.org/10.1145/267954.267957.

[15] M. F. Porter, An algorithm for suffix stripping. Program, 40(3)., 2006 https://doi.org/10.1108/00330330610681286

[16] Sh. Manish, A. Nitin, M. Bibhuti, Morphology based natural language processing tools for indian languages. In Proceedings of the 4th Annual Inter Research Institute Student Seminar in Computer Science, IIT, Kanpur, India, April. Citeseer.

[17] A. Goweder, H. Alhami, Tarik Rashed, and A Al-Musrati, A hybrid method for stemming Arabic text. Journal of computer Science.

[18] G. Adam, K. Asimakis, C. Bouras, V. Poulopoulos, An efficient mechanism for stemming and tagging: The case of Greek language. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6278 LNAI(PART 3)., 2010. https://doi.org/10.1007/978-3-642-15393-8_44

[19] D.Cutting, J.Kupiec, J.Pedersen, P.Sibun, A practical part-of-speech tagger, 1992. https://doi.org/10.3115/974499.974523

[20] B. Merialdo, Tagging English text with a probabilistic model. Computational Linguistics, 20(2)., 1994.

[21] M.Maimaiti, A.Wumaier, K.Abiderexiti, T.Yibulayin, Bidirectional long short-term memory network with a conditional random field layer for Uyghur part-of-speech tagging. Information (Switzerland), 8(4)., 2017. https://doi.org/10.3390/info8040157

[22] X. Qi Azragul, A. Yusup, “Website Phrasal Survey Based Modern Uighur Stem Extraction and Application Study”, Computer Applications and Software, vol.29, no.3, (2012), pp.32-34.