A GRAMMATICALLY AND STRUCTURALLY BASED PART OF SPEECH (POS) TAGGER FOR ARABIC LANGUAGE

Mohamed Taybe Elhadi[1] and Ramadan Sayad Alfared[2]

[1]Software Engineering Department,
Faculty of Information Technology, Zawia University, Az Zawiyah, Libya
[2]Computer Technology Department,
Faculty of Information Technology, Zawia University, Az Zawiyah, Libya

ABSTRACT

In this paper we report on an experimental syntactically and morphologically driven rule-based Arabic tagger. The tagger is developed using Arabic language grammatical rules and regulations. The tagger requires no pre-tagged text and is developed using a primitive set of lexicon items along with extensive grammatical and structural rules. It is tested and compared to Stanford tagger both in terms of accuracy and performance (speed). Obtained results are quite comparable to Stanford tagger performance with marginal difference favoring the developed tagger in accuracy with huge difference in terms of speed of execution. The newly developed tagger named MTE Tagger has been tested and evaluated. For the evaluation of its accuracy of tagging, a set of Arabic text was manually prepared and annotated. Compared to Stanford tagger, the MTE tagger performance was quite comparable. The developed tagger makes use of no pre-annotated datasets, except of some simple lexicon consisting of list of words representing closed word types like demonstrative nouns or pronouns or some particles. For the purpose of evaluation of the new tagger, it was run on multiple datasets and results were compared to those of Stanford tagger. In particular, both taggers (the MTE and the Stanford) were run on a set of 1226 sentences with close to 20,000 tokens that was human annotated and verified to serve as testbed. The results were very encouraging where in both test runs, the MTE tagger outperformed the Stanford tagger in terms of accuracy of 87.88% versus 86.67% for the Stanford tagger. In terms of speed of tagging and in comparison Stanford tagger, MTE Taggers' performance was on average 1:50. More improved accuracy is possible in future work as the set of rules are further optimized, integrated and more of Arabic language properties such as end of word discretization are used.

KEYWORDS

Part of Speech Tagging, Rule-based POS, Arabic Text Processing.

1. INTRODUCTION

Natural Language Understanding (NLU) and Text Processing with all of their related subfields are part of what is termed computational linguistics. Computational linguistics is the combination of computing and linguistics dealing with the automatic processing of natural language using artificial intelligence and machine learning methods [1,2,3].

Part of speech tagging (POS) is an important component of NLP and a prerequisite for many text-processing subfields and applications. It is concerned with assigning a label to language tokens representing most appropriate grammatical or morph-syntactical category. POS-tagging is usually the first step in linguistic analysis and a very important intermediate step in many applications such as machine translation, parsers, information retrieval, and spell-checkers-correctors [1]. POS has

DOI: 10.5121/ijnlc.2022.11502

very much matured for English and many European languages. Unlike English, however, Arabic language lacks NLP tools and resources including POS taggers and its prerequisite resources; the manually tagged corpora. When done on a large corpus, POS for instance, is a labor intensive and time-consuming task. As most POS processing is based on Markoff models which in turn are based on statistical models trained on annotated datasets. Obtaining a fast and quality tagging algorithms with high precision and accuracy requires a large manually annotated data. It is quite a disappointment for many language users, Arabic language users in particular, that there are very limited and hardly any freely available annotated data for training and evaluations. With the exception of very few tools, there is hardly any tools for Arabic part of speech tagging. Great many attempts have taken place to produce POS taggers for Arabic using non-statistical alternatives such as rule-based and machine learning methods.

This paper reports on an experimental syntactically and morphologically driven rule-based Arabic tagger developed using Arabic language grammatical rules and regulations without the use of pre-tagged corpora and without the involvement of any language experts except for the authors who are merely native Arabic speakers. It is developed using a primitive set of lexicon items along with extensive grammatical and structural rules. The tagger is tested and compared to currently used tagger in terms of accuracy and performance (speed). Obtained results are quite comparable and in favour of the newly developed tagger both in accuracy and certainly more in speed of tagging.

This paper is organized as follows: section 1 is an introduction, section 2 is a review of Arabic POS; Section 3 contains a description of corpora and datasets used for testing and evaluations; section 4 contains a detailed description and discussion on the new tagger (named MTE Tagger); Section 5 contains a detailed description of experiments conducted and results obtained along with discussions; finally, section 6 contains concluding remarks followed with list of reference used.

2. ARABIC LANGUAGE AND PART OF SPEECH TAGGING (POS)

Like many Semitic languages, Arabic language had a history that belongs to thousands of years ago [4]. The language is used as the language of journalism, media and education in the private sector as well as in public and governmental agencies. It is the medium of commutations for close to 400 Arabs in 21 states and large communities all over the globe. It is also of interest to many other nations who share the religious beliefs of Islam, as it is the language of the Holy Quran. The Department of Cultural Affairs in USA asserts that Arabic language is one of the worlds' important languages and the United Nations lists it as the sixth official language of the United Nations [5,6]. Modern Standard Arabic (MSA) is based on classical Arabic, on the wholly Quran and on Arabic literature. It is a written and read in a right to left fashion using a set of twenty-eight letters. Arabic has three numbers, singular, dual, and plural; two genders, feminine and masculine; and three grammatical cases, nominative, accusative, and genitive. Words are grouped as nouns, verbs, adjectives, adverbs, and particles [5,6,7].

Arabic has limited access to technology hindering research efforts in automation and utilization. With a relatively complex nature, Arabic itself poses quite a challenge for NLP [6]. Such challenges are further complicated by the existence of multiples of parallel dialects that are similar in certain aspects but different in others [5].

It is stated in [8] that Arabic is made of 58% nouns, 31% particles and 11% verbs with prepositions making up 14.1%, and 44.5% of all particles. Nouns in the genitive case make up 61.8% of the total, in the accusative 9.6% and in the nominative 18.5%. Nouns that occur after a preposition are more frequent than nominative and accusative nouns [9]. Such characteristics and challenges cause lots of ambiguities. Some such ambiguities are inevitable and constitute an inherent part of the language. They are rather considered advantageous by introducing flexibility and expressiveness from the perspective of authors particularly eloquent writers and poets.

There are other challenges and limitations that relate to POS that has to do with tag sets development and usage. Arabic' limited research work on standard tag-set, lack of resources, richly inflected nature and a complex morphological nature constitute the main reasons for such limited research. Arabic lacks manually tagged corpus to be used for training and evaluations making it hard to develop tools such as POS taggers using a statistical approach [10-13].

Tag sets, an important part of POS research, had to be developed and researched. Some of the researched and used tag sets include Brown tag set which contains 226 tags, LOB tag set which contains 135 tags which is based on the tag set that was used in Brown corpus and Penn Treebank. Other Arabic tag sets used in literature includes Khoja tag set with 177 detailed tags [14]. This set was used for their semi-automatic tagger system. In [15] a tag set of 55 tags were used for an HMM tagger. In [1] a tag set which contains 28 general tags and 161 detailed tags used by AMT tagger system [2].

2.1. Part of Speech Tagging (POS)

In natural languages, Arabic included, POS tagging is the process of assigning of words into specific and representative linguistic or grammatical type. It is the assignment of POS tags to a token taken from a sentence, a paragraph or simple set of words. POS tagging is an important requirement to NLP applications permitting the categorisation of words as adjectives, verbs, nouns, or a preposition. Knowing the verbs in a sentence, for instance can very well help us deciding on action(s) the sentence may contain and can help to indicate sentential meaning. With POS tagging we are also able to do chunking of sentences and aid in figuring out functional components such as Subject, Object and Verb. A number of different approaches are being used to do POS tagging with rule-based and stochastic methodology [16,17] as the most common methods.

2.1.1. Stochastic POS Models

Stochastic models use the probabilistic methods using first-order or second-order Markov models [18, 19]. They are based on building a trainable statistical language model and estimating parameters using previously tagged corpus. They make use of the idea that the probability of a word appearing with a specific tag and the probability that a tag is followed by another. Tree Tagger is and earlier example of such taggers that achieving an accuracy of up to 96.36% [19,20]. Stochastic systems require less work and cost than the rule-based approach and are considered more transporting of the language model to other languages especially provided that large manually tagged corpus is available. On the other hand, they suffer from unknown words that cannot be tagged and lack of annotated corpora in certain languages. Some of the stochastic based system in use include CLAWS (Constituent-Likelihood Automatic Word-Tagging System), PARTS system; and many other POS-taggers [21-26].

2.1.2. Rule-based POS Models

Rule-based taggers go back to the 1960-70's and use a set of linguistic rules during the tagging process. They are easy to maintain and provide an accurate and robust system [27-28], but are very difficult to build. Some of the well-known rule-based systems CGC (Computational Grammar Coder) [29,20], TAGGIT [32], TBL (Transformation-Based error-driven Learning) system [33] and Fidditch system [34] and many others [35].

2.1.3. Hybrid and Other Systems

A number of systems with high rate of accuracy were produced using a combination of more than one model. Examples include [35,36] for European languages with a reported accuracy of 98%.; POS-tagger for Hungarian language[12] and POS-tagger developed by [37,38]. Many other POS taggers are inspired from the Artificial Intelligence are developed including machine learning, memory based and neural networks [39-43]. Other methods used include Conditional Random Fields, Long Short-Term Memory (LSTM) and a variation on LSTM the bidirectional LSTMs (BiLSTMs), in which the learning algorithm is fed with the original data from the beginning to the end, and then from the end to beginning [44].

2.2. Arabic POS Tagging

An initial focal step in Arabic POS was the adaptation of Tree tagger for Arabic with tag set that covers 22 different languages including Arabic [45,46]. Later on, the Stanford POS tagger was introduced to become one of the few taggers that supported Arabic [47]. Khoja is an Arabic developed tagger that combines statistical and rule-based approach achieving an accuracy close to 90% [48]. Schmid reported on a language independent tagger based on decision trees [49,50]. Algerian, et al reported an accuracy of 91% using a small manually annotated lexicon [51]. Yousif, et al used the Support Vector Machines (SVM) and a tagged corpus reporting an accuracy of 99.99% [52]. Labidi reported on similar work [53] using augmented stately sliding-window [54] based on a database of nearly 50.000 Arabic terms.

Al Shamsi, et al reported on a semi-automatic hybrid tagger that used statistical method and morphological rules in the form of HMMs achieving an accuracy of 90% [55]. Othmane, et al reported on Automatic Arabic POS-Tagger which is a combination of statistical and rule-based techniques achieving an accuracy of 86 % [56]. Mohamed, et al implementing Arabic Brill's POS-tagger using a manually created corpus [57]. Kučera, et al created a rule-based tagger basing their work on automatic annotation output produced by the morphological analyser of Tim BuckeckWalter reported an accuracy of 96% [58]. An Arabic POS-tagger was developed using the support vector machine (SVM) method and LDC (Linguistic Data Consortium) [59]. An HMM tagger for Arabic language with an accuracy of 96% was reported on in [60]. In [61], a tagger was developed that used a rule-based and a memory-based learning achieving an accuracy of 86%.

In [62], structure of Arabic sentence was taken into account in developing an Arabic POS-tagger for un-vocalized text with an accuracy of 97%. In [63], a POS-tagger was developed using the rule-based and untagged raw partially-vocalized corpus making use of pattern-based, lexical, and contextual rules. The system an accuracy of 91%. In [64] used the genetic algorithm and a reduced tag set to develop an Arabic POS tagging. [65] considered the structure of Arabic sentence combining morphological analysis with Hidden Markov Models (HMMs) obtained a recognition rate of this tagger reached 96%. In [66, 67] developed what they termed whole word tagging and segmentation-based tagging. More research on Arabic POS tagging is performed. Most of such work uses the hidden Markov models [68-70]. Unfortunately, most of reported wok is private with limited availability.

3. CORPORA AND DATASETS

A corpus is normally looked at as a large collection of machine readable text that is accessible and searchable. It role is to provide POS systems with the needed linguistic knowledge that helps resolving the ambiguity. Well-known corpora for English language include Brown University [71], Bergen Corpus of London Teenage English (COLT) [72], BNC (British National Corpus) [73], Child Language Data Exchange System (CHILDES) [74], TOSCA Corpus [65] and Penn Treebank Corpus [75].

For Arabic language, however, there is no free corpus available. A few corpora were created for Arabic language including, but not limited to, LDC Arabic newswire corpus, Hayat newspaper corpus, An-Nahar Newspaper Text Corpus, Buckeck Walter Arabic Corpus, Nijmegen Corpus, Penn Arabic Treebank Corpus and Corpus of Contemporary Arabic [58,75,76,77].

As this work is based on linguistics rules and regulations of Arabic language, there was no need or use of corpora except for evaluation purposes. As will be discussed in section 5, five datasets referred-to in this work as CNN-UTF8, Basel-Dataset, Arabic Discretized Books, Quranic Text Dataset and Annotated Dataset were used for evaluation. They included two sets on news, one from Quran and one from books along with the annotated set.

4. THE MTE TAGGER: THE PROPOSED APPROACH

MTE Tagger is totally based on readily available primitive data lists and a complex set of linguistic rules both of which are highlighted next:

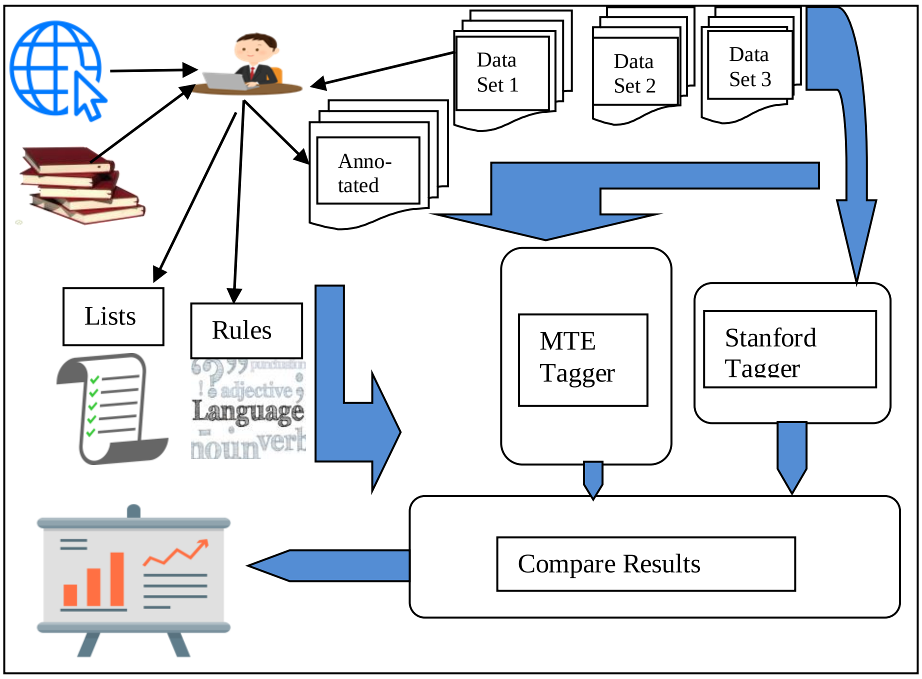

Figure 1. The over all process adopted.

Available word list of some known word types which are mostly collected from web, short and limited in size, hand formulated, closed word types of Arabic basic components such as pronouns, demonstrative nouns …etc. Few non-exhaustive lists are for things such as common names and geographical information like cities and countries …etc.

1. An initial data collection using available net resources and datasets is performed. This part is geared at minimal work to account for many of the closed Arabic word taxonomies and some limited but readily available list. It resulted in the lists as shown in the table (mostly incomplete but quite useful).

Table 1. Sample Closed POS Categories List

Days Weeks Months Seasons	Adjectives صفات	God Attributes صفات الله
Verb Cana-Thanaكان- ضن و اخواتها	Mausool Names اسم الموصول	Five Names الأسماء الخمس
Adjective Seq Numbers ارقام تراتبية	Punctuation ترقيم	Particles حروف
Place and Time Adverbs الضرف	Currencies عملات	تفضيل
Independent Pronounsضمائر منفصلة	Towns & countries مدن وبلدان	Exceptional Adverbs
Demonstrative Nouns اسم الاشارة	Colors الوان	Partials Rabat حروف الربط
IN Particle حروف الجر	Numbersارقام	MSA Verb افعال
Particle Ena'aان واخواتها	God Names الأسماء الحسنى	People Names أسماء الأشخاص

2. A second set of very compacted grammatical functions or rule representing the Arabic language rules for many of syntactical and structural compositions. Table 2 lists some names of some of such rules. The tagger works sentence by sentence by first searching the set of lists (primitive lexicon) to decide on the category of any word in the sentence. Next it will reiterate through the set of grammatical and structural rules and will re-confirm existing categories and fix new ones. At the end, a sentence may have some missing categories for some words. A number of experiments were tried but the issue still open for improvements.

Table 2. Gramatical, Morphological and Strctural Procedures (Rules)

Ckeck Eshar Plus	Ckeck FirstPOS	Ckeck IfAdaadTratubiah
Ckeck SplDarfI	Ckeck Tarkim PunI	Ckeck TimesI Ckeck
Ckeck Spl Words I	Ckeck CDI	HrfJarAndAftNNContextI
Ckeck MudenBuldanI	EnaaGroupCTX	tamizeAlafadd
Ckeck MaousoolNameCtxI	Ckeck TafdealCtx1I	Emma
Ckeck IfParticlesNotJar	Ckeck IstithnaCtxI	EthandEthaPlus
Ckeck Names	Ataf GroupCTX	Ckeck Lema Ctx
Ckeck Damer MunfaselI	Ckeck StartWaAlifCtxI	BaseRuleSet1
Ckeck WITHCtxI	Ckeck MaaaCt	Ckeck IfEThCtx
CanaZanaGroupCtx	Ckeck MustathnaI	Ckeck AlfLamCtxI
LaNafiaWaNahia	Ckeck AlfFariqaCtxI	Ckeck LELI
Ckeck KADCtxI	ckeckSaYAVBCtx	setJJ

5. EXPERIMENTS, EVALUATIONS, AND DISCUSSIONS

The evaluations process consists of the following experiments with results as shown in Table 3. Two sets of experiments were performed. The first set is made of four runs on four different un-annotated data sets to compare performance (Accuracy and Timing) of the new tagger to that of Stanford Tagger.

Table 3. Accuracy and Timings results comparison.

	Accuracy	**Stanford Tagger**	**MTE Tagger**	**Speed**
Data/Sets Experiments	**MTE / STF**	**Timing Mints**	**Timing Mints**	**%**
CNN-UTF8	74.23	2954.5	4.4	0.059
Basel-Dataset	75,2	714.13	0.245	0.0003
Arabic Discretized Books	75.45	6052.57	62.54	0.0103
Quranic Text Dataset	65.45	3252.73	4.1	0.0013
Average	**72.58**	**3243;48**	**71.29**	**0.022**
Annotated Dataset	**87.88 / 86.67**	**0.022**	**0.020**	**.91**

The second set of experiments are based on a small selected dataset that is manually annotated. The two taggers are both run on the data set and accuracy of tagging and speed of performance are noted and compared. Accuracy is a representation of the number of rightly tagged tokens while performance is the speed of tagging. Due to the expectation that rule-based systems tend to be much faster and robust, the measurements take are only indicative and lack features of a well-controlled experiments.

5.1. MTE Tagger vs. Stanford Tagger on Unannotated Datasets

This is a set of four experiments each is conducted on a different set of data.

5.1.1. Experiment-1: MTE Tagger vs. Stanford tagger on CNN-UTF8

This a set of 5070 files of news articles taken from CNN covering business, entertainment, middle east, science, technology, sports and world news [1]. The whole set contains a total of 141,4021 words. The obtained results of comparing MTE tagger to Stanford tagger showed and overall accuracy of 74.20 %. That is the percentage of overlap between the taggers is three quarters and differed on one quarter. In terms of timings, the results were 48 minutes and 44.5 seconds versus only 4.4 seconds respectively. That is the time taken by MTE Tagger is only 0.0015% of that taken by Stanford tagger.

5.1.2. Experiment-2: MTE Tagger vs. Stanford on Basel-Dataset

This is a set of 1000 files that are hand compiled on 4 different subjects, namely science, politics, arts, sports and economics [78]. The set is made of a total of 159,442 words. The obtained results of comparing MTE tagger to Stanford tagger show and overall accuracy of 75.12%. Again numbers are similar to the previous set. In terms of timings, the results were 11 minutes and 54.13 seconds versus 0.245 seconds which is only 0.00034%

5.1.3. Experiment 3: MTE Tagger vs. Stanford on Arabic Discretized Books (Tashkeela)

This a set of 20 files with each file containing the text of a whole book [79]. The text is discretized. The set is made of a total of 298,416 words. The obtained results of comparing MTE tagger to Stanford tagger show and overall accuracy of 75.43%. Again numbers are similar to the previous two sets. In terms of timings, the results were 1 hour, 40 minutes and 52.57 seconds versus 1 minute and 1.54 seconds which is .01%. This albeit higher than the previous cases. It is probably due to diacritics removal.

5.1.4. Experiment 4: Quranic Text Dataset

This a single file containing 214 chapters of the whole Quran. The set is made of 77,289 words. The obtained results of comparing MTE tagger to Stanford tagger show and overall accuracy of 65.79%. Even though this lower than previous experiments results, still however within an acceptable range. In terms of timings, the results were 54 minutes and 12.73 seconds versus 4.1 seconds which is only 0.0013%. On average, there percentage accuracy is 88.89% and the speed is 2.2% faster in favor of MTE Tagger.

5.2. MTE Tagger vs. Stanford Tagger on Annotated Dataset:

This experiment is based on manually tagged dataset. The data set consists of a total of 17,485 words taken from set 1. The objective is comparing the MTE tagger performs to that of Stanford tagger. The results obtained were 87.89 % for MTE Tagger and 86.67% for the Stanford tagger. In terms of timings, the results showed that MTE took on average only 2.2% of the time taken by Stanford tagger.

5.3. Discussions

From obtained results, we can see that the first set of experiments aimed at looking at the using of Stanford tagger on four datasets with variable tokens and context. The data set included 2 sets on news, one from Quran and one from books.

To validate the utility of Stanford and then to compare the results obtained to MTE tagger, the experiments were clearly that the MTE tagger did not compare well to the Stanford tagger with an average of 72.64%. That is to say they only agree 72.64%. This prompted us to study the data and see the differences and agreements. It was clear that the difference was a result of disagreements for which many cases Stanford failed to tag correctly and vice versa.

Obtained results prompted us to annotate part of the datasets to use for evaluation. We studied the results and confirmed the correct and fixed the incorrect for a set of 1226 sentences.

The manually annotated set was developed and compared to the two taggers. Obviously better accuracies were made with MTE tagger having a marginally higher accuracy.

The obtained results were very encouraging and further refinement of the tagger to include more complete lexicon and more rules using furthers linguistic properties like vocalization will certainly make the tagger perform better. As far as time performance, much better numbers where obtained with MTE taking only 2.2% compared to STF. It is expected that Rule-based are to be much faster, but we were surprised by their results. Table 5 shows a sentence taken from the results.

Table 4. A sentence example: Made of 27 Tokens. Taggers match on 19 and mismatch on 8

Word	**MTE**	**STF**	**Agree**	**Verify**	**0/1**	**Word**	**MTE**	**STF**	**Agree**	**Verify**	**0/1**
وبينت	VBD	VBD	Agree	Both Right	1	معهم	NN	JJ	Disagree	STF Wrong	1-0
الشرطة	DTNN	DTNN	Agree	Right	1	ما	WP	WP	Agree	are Right	1
ان	IN	IN	Agree	Right	1	يبدو	VBP	VBP	Agree	are Right	1
صلة	NN	NN	Agree	Right	1	لافتة	NN	JJ	Disagree	are wrong	0-0
المهاجم	DTNN	DTNN	Agree	Right	1	المحققين	DTNNS	DTNNS	Agree	Right	1
تميل	VBP	VBP	Agree	Right	1	تواصلوا	VBD	VBD	Agree	are Right	1
الي	IN	IN	Agree	Right	1	مع	IN	NN	Disagree	STF Wrong	1-0
انه	NN	NN	Agree	Right	1	زوجته	NN	NN	Agree	Right	1

متاثر	NN	NN	Agree	Right	1	دون	RB	NN	Disagree	STF Wrong	1-0
بالتنظيم	NN	NNP	Agree	Right	1	تقديم	NN	NN	Agree	Right	1
عوضا	NN	NN	Agree	Right	1	تفاصيل	VBP	NN	Disagree	MTE Wrong	0-1
عن	IN	IN	Agree	Right	1	اضافية	-----	JJ	Disagree	MTE Failed	0-1
علي	IN	IN	Agree	Wrong	0	تفاصيل	VBP	NN	Disagree	MTE Wrong	0-1
تواصل	VBP	NN	Disagree	MTT Wrong	0-1	اضافية	-----	JJ	Disagree	MTE Failed	0-1
مباشر	NN	JJ	Disagree	MTT Wrong	0-1	-	-	-	-	-	-

The calculated result is 19/27. This signifies that both the calculated percentage overlap can still be made more accurate.

Table 5. Overal missed percentage for the different POS catagories

Tag	**DTNNPS**	**DTNNP**	**DTJJR**	**DT**	**CD**	**IN**	**NNS**
Wrong	0	0		0	4	34	17
Right	50	35	16	126	647	1519	608
%	**0**	**0**	**0**	**0**	**0.62**	**2.24**	**2.8**
Tag	WP	CC	DTNNS	RP	NNP	JJR	RB
Wrong	7	8	13	4	15	3	32
Right	243	265	379	99	319	53	462
%	**2.88**	**3.02**	**3.43**	**4.04**	**4.7**	**5.66**	**6.93**
Tag	PRP	DTNN	VBD	NN	JJ	DTJJ	VBP
Wrong	6	532	88	1048	10	102	596
Right	84	5432	674	7014	56	537	916
%	**7.14**	**9.79**	**13.06**	**14.94**	**17.86**	**18.99**	**65.07**
Tag	Nouns	Verbs	Adjectives	**<<For**	**totals**		
Wrong	1625	684	112				
Right	13837	1590	593				
%	**11.74**	**43.02**	**18.89**				

Looking at the overall success of tagging we could see that Adjectives (JJ) are the least accurate in MTE and better rules will still have to be invented to improve the classification of JJs.

6. CONCLUSIONS

In this paper we report on an experimental syntactically and morphologically driven rule-based Arabic tagger. The tagger is developed using Arabic language grammatical rules and regulations. The tagger requires no pre-tagged text and is developed using a primitive set of lexicon items along with extensive grammatical and structural rules. It is tested and compared to Stanford tagger both in terms of accuracy and performance (speed). Obtained results are quite comparable to Stanford tagger performance with marginal difference favoring the developed tagger in accurate and huge difference in terms of performance. The newly developed tagger name MTE Tagger has been tested and evaluated and was able to obtain an accuracy of 85% versus 82% for the Stanford tagger.

The developed tagger makes use of no pre-annotated datasets, except of some simple lexicon consisting of list of words representing closed word types like demonstrative nouns or pronouns list or some particles. For the purpose of evaluation of the new tagger, it was run on multiple datasets and results were compared to those of Stanford tagger. In particular, both taggers (the MTE and the Stanford) were run on a set of 1226 sentences with close to 20,000 tokens that was human annotated and verified to serve as testbed. The results were very encouraging in both test runs the MTE tagger outperformed the Stanford tagger with accuracies in the range of 87.88% versus 86.67% for the Stanford tagger. In terms of efficiency (speed of tagging) the MTE to Sanford tagger 1:50.

Better accuracy is expected as the set of rules are optimized and other Arabic language properties such as end of word discretization are used.

REFERENCES

[1] Abumalloh RA, Al-Sarhan HM, Ibrahim O, Abu-Ulbeh W. Arabic part-of-speech tagging. Journal of Soft Computing and Decision Support Systems. 2016 Feb 25;3(2):45-52.

[2] Das BR, Sahoo S, Panda CS, Patnaik S. Part of speech tagging in odia using support vector machine. Procedia Computer Science. 2015 Jan 1;48:507-12.

[3] Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, Ess-Dykema CV, Meteer M. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational linguistics. 2000 Sep 1;26(3):339-73.

[4] Seikaly ZA. The Arabic Language: The Glue that Binds the Arab World. AMIDEAST, America-Mideast Educational and Training Services, Inc. 2007

[5] Elhadi MT, Alfared RA. Adopting Arabic Taggers to Annotate a Libyan Dialect Text with a Pre-Tagging Processing and Term Substitutions. International Journal of Science and Technology, Specail Edition. Feb 2022.

[6] Elhadi MT. Arabic News Articles Classification Using Vectorized-Cosine Based on Seed Documents. Journal of Advances in Computer Engineering and Technology. 2019 May 1;5(2):117-28.

[7] Agirre E, Edmonds P, editors. Word sense disambiguation: Algorithms and applications. Springer Science & Business Media; 2007 Nov 16.

[8] al-Khūlī, M.A. al-Tarākīb al-Shāiʻah fī al-Lughah al-ʻArabīyah: Dirasah Iḥṣāʼīyah .ʻAmmām: Dār al-Falāḥ lil-Nashr wa-al-Tawzīʻ. (1998).

[9] Husni R, Zaher A. Working with Arabic prepositions: Structures and functions. Routledge; 2020 Mar 2.

[10] Marsi E, Van Den Bosch A, Soudi A. Memory-based morphological analysis generation and part-of-speech tagging of Arabic. InProceedings of the ACL workshop on computational approaches to semitic languages 2005 Jun (pp. 1-8).

[11] Stokoe C, Oakes MP, Tait J. Word sense disambiguation in information retrieval revisited. InProceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval 2003 Jul 28 (pp. 159-166).

[12] Atkins BT. Tools for computer-aided corpus lexicography: the Hector project. Acta Linguistica Hungarica. 1992 Jan 1;41(1/4):5-71.

[13] Jacquemin B, Brun C, Roux C. Enriching a text by semantic disambiguation for information extraction. arXiv preprint cs/0506048. 2005 Jun 12.

[14] Chiche A, Yitagesu B. Part of speech tagging: a systematic review of deep learning and machine learning approaches. Journal of Big Data. 2022 Dec;9(1):1-25.

[15] Schneider, S. The biggest data challenges that you might not even know you have, https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

[16] Masand B, Linoff G, Waltz D. Classifying news stories using memory based reasoning. InProceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval 1992 Jun 1 (pp. 59-65).

[17] Zaghouani W. Critical survey of the freely available Arabic corpora. arXiv preprint arXiv:1702.07835. 2017 Feb 25.

[18] Bahl, L.R., Mercer, R.L.: Part-of-speech assignment by a statistical decision algorithm. In: Proceedings of the IEEE International Symposium on Information Theory, pp. 88-89. IEEE Computer Society Press, Los Alamitos (1976)

[19] Sanger N. The computational analysis of English: A Corpus-based approach: R. Garside; G. Leech; and G. Sampson (Eds.). Longman, London and New York (1987).

[20] Marshall I. Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus. Computers and the Humanities. 1983 Sep 1:139-50.

[21] Church KW. A stochastic parts program and noun phrase parser for unrestricted text. InInternational Conference on Acoustics, Speech, and Signal Processing, 1989 May 23 (pp. 695-698). IEEE.

[22] Cutting D, Kupiec J, Pedersen J, Sibun P. A practical part-of-speech tagger. InThird conference on applied natural language processing 1992 Mar (pp. 133-140).

[23] Kupiec J. Robust part-of-speech tagging using a hidden Markov model. Computer speech & language. 1992 Jul 1;6(3):225-42.
[24] Weischedel R, Schwartz R, Palmucci J, Meteer M, Ramshaw L. Coping with ambiguity and unknown words through probabilistic models. Using Large Corpora. 1994:323-6.
[25] Merialdo B. Tagging English text with a probabilistic model. Computational linguistics. 1994;20(2):155-71.
[26] Khoja S. APT: An automatic Arabic part-of-speech tagger (Doctoral dissertation, Lancaster University). 2003
[27] Alqrainy S. A morphological-syntactical analysis approach for Arabic textual tagging (Doctoral dissertation, De Montfort University). 2008
[28] McEnery AM. Computational Linguistics: A handbook and toolbox for natural language processing. Coronet Books Incorporated; 1992.
[29] Harris Z. String analysis of language structure. Mouton and Co., The Hague. 1962.
[30] Klein S, Simmons RF. A computational approach to grammatical coding of English words. Journal of the ACM (JACM). 1963 Jul 1;10(3):334-47.
[31] Greene BB, Rubin GM. Automatic grammatical tagging of English. Department of Linguistics, Brown University; 1971.
[32] Brill E. A simple rule-based part of speech tagger. PENNSYLVANIA UNIV PHILADELPHIA DEPT OF COMPUTER AND INFORMATION SCIENCE; 1992 Jan 1.
[33] Abney S. Part-of-speech tagging and partial parsing. InCorpus-based methods in language and speech processing 1997 (pp. 118-136). Springer, Dordrecht.
[34] Chanod JP, Tapanainen P. Tagging French--comparing a statistical and a constraint-based method. arXiv preprint cmp-lg/9503003. 1995 Mar 2..
[35] Volk M, Schneider G. Comparing a statistical and a rule-based tagger for German. arXiv preprint cs/9811016. 1998 Nov 11.
[36] Schmid H. Part-of-speech tagging with neural networks. arXiv preprint cmp-lg/9410018. 1994 Oct 24.
[37] Perez-Ortiz JA, Forcada ML. Part-of-speech tagging with recurrent neural networks. InIJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222) 2001 Jul 15 (Vol. 3, pp. 1588-1592). IEEE.
[38] Bosch AV, Marsi E, Soudi A. Memory-based morphological analysis and part-of-speech tagging of Arabic. In Arabic Computational Morphology 2007 (pp. 201-217). Springer, Dordrecht.
[39] Cover T, Hart P. Nearest neighbor pattern classification. IEEE transactions on information theory. 1967 Jan;13(1):21-7.
[40] Zavrel J, Daelemans W. Recent advances in memory-based part-of-speech tagging. In VI Simposio Internacional de Comunicacion Social 1999 (pp. 590-597).),
[41] Ramani D. A short survey on memory based reinforcement learning. arXiv preprint arXiv:1904.06736. 2019 Apr 14.
[42] Angle S, Mishra P, Sharma DM. Automated Error Correction and Validation for POS Tagging of Hindi. In Proceedings of the 32nd Pacific Asia Conference on Language, Information and computation 2018.
[43] Kanakaraddi SG, Nandyal SS. Survey on parts of speech tagger techniques. In2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) 2018 Mar 1 (pp. 1-6). IEEE.
[44] Katz G, Diab M, editors. Introduction to the Special Issue on Arabic Computational Linguistics. ACM Transactions on Asian Language Information Processing (TALIP). 2011 Mar 1;10(1):1-4.
[45] Toutanvoa K, Manning CD. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora 2000 Oct (pp. 63-70)
[46] Manning CD. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In International conference on intelligent text processing and computational linguistics 2011 Feb 20 (pp. 171-189). Springer, Berlin, Heidelberg.
[47] Brill E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational linguistics. 1995;21(4):543-65.
[48] Khoja S. APT: Arabic part-of-speech tagger. InProceedings of the Student Workshop at NAACL 2001 Jun (pp. 20-25).

[49] Schmid H. TreeTagger-a language independent part-of-speech tagger. http://www. ims. uni-stuttgart. de/projekte/corplex/TreeTagger/. 1994

[50] Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH. Top 10 algorithms in data mining. Knowledge and information systems. 2008 Jan;14(1):1-37.

[51] Algrainy S, AlSerhan HM, Ayesh A. Pattern-based algorithm for part-of-speech tagging. In International Conference on Computer Engineering and Systems, ICCES 2008 (pp. 119-124).

[52] Yousif JH, Sembok TM. Arabic part-of-speech tagger-based Support Vectors Machines. In2008 International Symposium on Information Technology 2008 Aug 26 (Vol. 3, pp. 1-7). IEEE.

[53] Labidi M. New Combined Method to Improve Arabic POS Tagging. Journal of Autonomous Intelligence. 2019 Jan 8;1(2):23-8.

[54] Sánchez-Villamil E, Forcada ML, Carrasco RC. Unsupervised training of a finite-state sliding-window part-of-speech tagger. InInternational Conference on Natural Language Processing (in Spain) 2004 Oct 20 (pp. 454-463). Springer, Berlin, Heidelberg.

[55] Elhadj YO. Statistical part-of-speech tagger for traditional Arabic texts. Journal of computer science. 2009;5(11):794.

[56] Diab M, Hacioglu K, Jurafsky D. Automatic tagging of Arabic text: From raw text to base phrase chunks. InProceedings of HLT-NAACL 2004: Short papers 2004 (pp. 149-152).Banko and Moore (2004)

[57] Al Shamsi F, Guessoum A. A hidden Markov model-based POS tagger for Arabic. InProceeding of the 8th international conference on the statistical analysis of textual data, France 2006 (pp. 31-42).

[58] Alqrainy S, Muaidi H, Alkoffash MS. Context-free grammar analysis for Arabic sentences. International Journal of Computer Applications. 2012 Jan 1;53(3).

[59] Othmane CZ, Fraj FB, Limam I. POS-tagging Arabic texts: A novel approach based on ant colony. Natural Language Engineering. 2017 May;23(3):419-39.

[60] Mohamed E, Kübler S. Arabic part of speech tagging. InProceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10) 2010 May.

[61] Alqrainy S. A morphological-syntactical analysis approach for Arabic textual tagging (Doctoral dissertation, De Montfort University).2008

[62] Kučera H, Francis WN. Computational analysis of present-day American English. Dartmouth; 1967.Stenstrom et al., 2002)

[63] Al-Taani AT, Al-Rub SA. A rule-based approach for tagging non-vocalized Arabic words. Int. Arab J. Inf. Technol.. 2009 Jul 1;6(3):320-8..

[64] AbuZeina D, Al-Khatib W, Elshafei M, Al-Muhtaseb H. Toward enhanced Arabic speech recognition using part of speech tagging. International Journal of Speech Technology. 2011 Dec;14(4):419-26.

[65] Tlili-Guiassa Y. Hybrid method for tagging Arabic text. Journal of Computer science. 2006;2(3):245-8.

[66] Hadni M, Ouatik SA, Lachkar A, Meknassi M. Hybrid part-of-speech tagger for non-vocalized Arabic text. Int. J. Nat. Lang. Comput. 2013 Dec;2(6):1-5.

[67] Calciu RH. Semantic change in the age of corpus linguistics. Journal of Humanistic and Social Studies. 2012;3(1):45-58. (MacWhinney and Snow, 1984),

[68] Ali BB, Jarray F. Genetic approach for Arabic part of speech tagging. arXiv preprint arXiv:1307.3489. 2013 Jul 11.

[69] El Hadj Y, Al-Sughayeir I, Al-Ansari A. Arabic part-of-speech tagging using the sentence structure. InProceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt 2009 Apr (pp. 241-5).

[70] Kübler S, Mohamed E. Part of speech tagging for Arabic. Natural Language Engineering. 2012 Oct;18(4):521-48.

[71] Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."; 2009 Jun 12.

[72] Stenström AB, Breivik LE. The Bergen corpus of London teenager language (COLT). ICAME journal. 1993 Apr;17:128.

[73] MacWhinney B. Child Language Data Exchange System. Transcript Analysis. 1984 Aug;1(1):2-18.

[74] Aarts J, Van Halteren H, Oostdijk N. The linguistic annotation of corpora: The TOSCA analysis system. International journal of corpus linguistics. 1998 Jan 1;3(2):189-210.

[75] Taylor A, Marcus M, Santorini B. The Penn treebank: an overview. Treebanks. 2003:5-22.

[76] Motaz K. Saad and Wesam Ashour, "OSAC: Open-Source Arabic Corpus", 6th ArchEng International Symposiums, EEECS'10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, 2010.
[77] Sklyarova DD. The History of Corpus Linguistics Development.
[78] Bani-Ismail, B, Al-Rababah, K, Shatnawi, S., The effect of full word, stem, and root as index-term on Arabic information retrieval, Global Journal of Computer Science and Technology, 2011
[79] Zerrouki T, Balla A. Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. Data in brief. 2017 Apr 1;11:147-51.