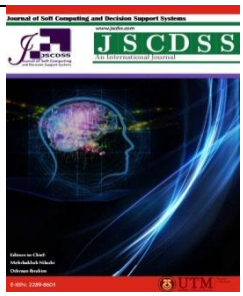
An International Journal
JSCDSS
http://www.jscdss.com
Vol.3 No.2 April 2016: 45-52
Article history:
Accepted 23 February 2016
Published online 23 February 2016

Journal of Soft Computing and Decision Support Systems



E-ISSN: 2289-8603

Arabic Part-of-Speech Tagging

Rabab Ali Abumalloh [a,*], Hassan Maudi Al-Sarhan [b], Othman Bin Ibrahim [c], Waheeb Abu-Ulbeh [c]
[a] University of Dammam, Department of Computer Science, Dammam, Saudia Arabia
[b] Ajloun National University, Faculty of Information Technology, Ajloun, Jordan
[c] Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia

*** Corresponding author email address**: ramolloh@uod.edu.sa

Abstract

The study described in this paper belongs to the area of computational linguistics. Computational linguistics is a field of artificial intelligence dealing with the logical modeling of natural language from a computational perspective. It unites two areas that are quite different in appearance, computer science and natural languages. Computational linguistics might be considered as a synonym of automatic processing of natural language, since the main task of computational linguistics is just the construction of computer programs to process words and texts in natural language. There are many areas that may be considered as properly included within the discipline of computational linguistics. One of these areas is part-of-speech tagging (POS-tagging). POS-tagging is considered as a process for automatically assigning the proper grammatical tag to each word of a written text according to its appearance on the text. Thus, the task of POS-tagging is attaching appropriate grammatical or morpho-syntactical category labels to each word, token, symbol, abbreviation and even punctuation mark in a corpus. POS-tagging is usually the first step in linguistic analysis. Also, it is very important intermediate step to build many natural language processing applications. It could be used in spell checking and correcting systems, speech recognition systems, information retrieval systems and text-to-speech synthesis systems.

Keywords: Tagging, Natural language processing, Arabic language

1. Introduction

Part-of-Speech tagging belongs to the field of computational linguistics. Computational linguistics is a discipline of artificial intelligence that deals with the logical modelling of natural language from a computational perspective (Das et al., 2015). It combines two fields: computer science and natural language processing (NLP). Computational linguistics focuses on proving linguistics theories using computer (Muaidi and Ayesh, 2006). These days the computational linguistics is concerned as one of the hottest topic research areas. Some of the areas that are studied by computational linguistics are: design of machine translation, design of parsers, design of spell-checkers and correctors.

In literature, there are many definitions for the term Part-of-Speech tagging". The well-known definition is the one that is introduced by Jurafsky and Martin (2000) in their valuable book: "Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition" This definition is as follows: "Part-of-speech tagging (or just tagging for short) is the process of assigning a part-of-speech or other syntactic class marker to each word in a corpus". POS-tagging is considered as a base stage for many of Natural Language Processing (NLP) applications. Some of these applications are:

i. Parsing System: For any parsing system there is a need for lexical information to be added to the words, this process is done using POS tagging.
i. Information Retrieval: Enhancing information retrieval applications using POS tagging process can be done by removing the lexical ambiguity and adding the syntactic class of the words.
ii. Building Dictionaries: The tagged text will be very useful by providing the required information needed for building dictionaries.
iii. Speech Synthesis Systems: Increasing the accuracy of speech recognition systems by producing more natural pronunciations in the speech synthesis systems.
iv. Word Processing: If the class of the misspelled word is known, then the process of correcting the word will be easier.

Computer corpus linguistics concerned with using large quantities of texts in machine-readable form. In the study of linguistic phenomena a corpus is defined by Leech (1992) as: "A Corpus is a large collection of natural language material stored in machine readable form that can be easily accessed, automatically searched, manipulated, copied and

transferred". Corpus provides POS systems with the needed linguistic knowledge that helps resolving the ambiguity in the language without the need for strong linguistic skills. Human effort for fully utilizing the corpora (plural of corpus) was an impossible job that required hundreds of individuals and involved errors, but the advanced potentials in the computer age makes this job much easier.

Brown corpus is considered as the first printed American English corpus that was created as an electronic corpus at Brown University by Kucera and Francis (1967). The work of this corpus had a big role in the development of computer software. Brown corpus contained about million words. After Brown corpus, hundreds of corpora ware created for English language for spoken and written texts. These corpora played an important role in compilation of dictionaries in the last few years. Some of the well-known corpora for English language are: Bergen Corpus of London Teenage English (COLT) which was compiled in 1993 (Stenstrom et al., 2002)**,** BNC (British National Corpus) which was created by Oxford University Press in the 1980s-1990s (Calciu, 2012), Child Language Data Exchange System (CHILDES) (MacWhinney and Snow, 1984), TOSCA Corpus (Aarts et al., 1998) and Penn Treebank Corpus (Taylor, 2003).

For Arabic language no free corpus was available. A few corpora were created for Arabic language. Some of the well-known corpora for Arabic language are: LDC Arabic newswire corpus (Alqrainy, 2008), Hayat newspaper corpus (2002), An-Nahar Newspaper Text Corpus from 1995 to 2000 (Maamouri et al., 2004), Buckwalter Arabic Corpus (2002), Nijmegen Corpus in 1996 (Alqrainy, 2008), Penn Arabic Treebank Corpus in 2001(Maamouri et al., 2004) and Corpus of Contemporary Arabic (CCA) (Al-Sulaiti et al., 2006).

The tagset is considered as all the possible tags that could be labelled to the words during the tagging process (Alqrainy, 2008). It is considered as a basic component in any POS-tagger system. There are a small number of popular tagsets for English language. All of them have been developed for many POS-tagger systems. Some of these tagsets are: Brown tagset which contains 226 tags that ware used to tag Brown corpus (Kucera and Francis, 1967), LOB tagset which contains 135 tags that ware used to tag LOB corpus and it was based on the tagset that was used in Brown corpus (Francis and Kucera, 1979) and Penn Treebank tagset which contains 36 tags and ware used to tag Penn Treebank corpus from 1989 to1996 (Taylor, 2003).

For Arabic language, tagsets have been developed for many POS-tagger systems. Some of these tagsets are: Khoja tagset which contains 177 detailed tags that were used for her APT tagger system (Khoja, 2003), El-Kareh and Al-Ansary tagset which contains 72 tags that were used for their semi-automatic tagger system (2000), Al-Shamsi and Guessom (2006) tagset which contains 55 tags that were used for their HMM tagger system and Al-Qrainy tagset (2008) which contains 28 general tags and 161 detailed tags that were used for his AMT tagger system.

Arabic language belongs to the family of Semitic languages which includes Tigrinya, Tigre, Amharic, Modern Hebrew, Syriac, some Aramaic dialects and Maltese (Seikaly, 2007). Semitic languages had a history belongs to thousands of years long ago (Seikaly, 2007). Moreover, most of these languages have died but Arabic language still has its dazzling and popularity among all Arabs and Muslims. Arabic language gave Arab word its unique identity. The advent of Islam and AL-Qur'an in the seventh century gave the Arabic language a great religious role. This is a result of the holy book which was revealed in Arabic language to the prophet Muhammad and it must be read in Arabic in order to understand the great message of AL-Islam. The presence of AL-Islam had extended the importance of the Arabic to include over one billion Muslims over the word. Arabic language needed by all Muslims around the world. They need Arabic when they are reading AL-Qur'an, in their praying even in their understanding of their religion. AL-Qur'an reflects the beauty and the perfection of Arabic language.

Scientists in Western Europe wanted to take advantage of the sciences developed by Arabs. Many books were translated from Arabic language to other languages. Because of the huge development in the technology many new foreign words were invented and accepted in Arabic language. Many foreign words related to modern technology entered the Arabic language like television, radio and telephone. These new foreign words were used to derive new words from it.

2. Literature review

Arabic language differs than Indo-European- languages, so taggers that have been developed for that languages may not be suitable for Arabic language. In order to illustrate these differences, a revision and a summarization of the development of POS-tagging systems over the years for English language are presented. The history of POS taggers that have been developed for Arabic language are reviewed and discussed.

In literature, there are two main methodologies for automatic POS tagging: rule-based methodology and stochastic (probabilistic) methodology (Jiyad, 2006). Most of POS-tagging systems have been implemented using these two methodologies. Some of the existing systems combined the two methodologies to produce a hybrid one which uses both methodologies, and some other systems use other approaches. The summary of the commonly approved methods for the POS taggers is presented in the following sections

2.1 Rule-Based Approach

The rule-based approach is the earliest approach that was used for automated POS-tagging (Khoja, 2003). The starting using of the rule-based approach goes back to the 1960's and 1970's. The rule-based approach tries to use a set of linguistic rules during the tagging process (Alqrainy,

An International Journal JSCDSS E-ISSN: 2289-8603

2008; Mc Enery, 1992). The number of rules that is used in the tagging process could differ from hundreds to thousands, so a huge work and cost is required. The rule-based technique is not very robust and difficult to build. Because knowledge representation in the rule-based approach is in the form of rules, it does not need a huge amount of stored information .This approach is considered to be easy to maintain and provides an accurate systems.
Some of the well-known rule-based systems are: POS tagger which was developed by Harris (1962), CGC (Computational Grammar Coder) system was developed by Klein and Simmons (1963), TAGGIT system that was developed by Greene and Rubin (1971), TBL (Transformation-Based error-driven Learning) system which was developed by Brill (1992) and Fidditch system that was developed by Hindle (1983).

2.2 Stochastic Approach

The stochastic approach is also known as statistical or probabilistic approach. It is based on building a statistical language models (trainable models) and estimating parameters using previously tagged corpus.

The statistical language model is built by collecting statistics from existing corpora. Some of these statistics parameters are:

i. Lexical Probability: The probability that a certain word appears with a certain tag.
ii. Contextual Probability: The probability that a tag followed by another.

Stochastic approach requires less work and cost than the rule-based approach. It is considered as the most popular approach of POS-tagging. It is also considered to be more transporting of the language model to other languages especially when a huge manually tagged corpus is available. Probabilities can be calculated automatically from the corpus. The main problem in statistical method that unknown words cannot be tagged using this approach. Hidden Marcov Models (HMMs) is an example of statistical approach. HMMs describe a stochastic tagging algorithm that is concerned with modelling a sequence of tags in a sentence. It is called hidden, because the sequence of tags is hidden from the observer of the text. In this approach, a sequence of words that forms a sentence is given and the tagger determines the set of tags these words belong to.

The following list shows some of the systems that ware implemented based on the stochastic approach:

i. POS-tagger which was developed by Bahl and Mercer (1976).
ii. CLAWS (Constituent-Likelihood Automatic Word-Tagging System) that was developed by Garside (1987).
iii. System that was developed by Marshal et al. (1983).
iv. PARTS system that was developed by Church (1988).
v. POS-tagger that was developed by Cutting (1992).
vi. POS-tagger that was developed by Kupiec (1992).
vii. POS-tagger that was developed by Weischedel (1993).
viii. POS-tagger that was developed by Merialdo (1994).

2.3 Hybrid Approach

Hybrid taggers approach combines both rule-based and stochastic approach methods. This approach achieved high rate of accuracy. Tapanainen and Voultilainen (1994) developed a tagger for French language that used both techniques separately and achieved an accuracy of 98 %.

Some of the systems that ware implemented based on the hybrid approach are: POS-tagger developed by Chanod and Tapanainen for French language (1995), POS-tagger developed by Kuba et al. (2004) for Hungarian language and POS-tagger developed by Schneider and Volk (1998).

2.4 Other Approaches

Some of these approaches ware inspired from the Artificial Intelligence field such as machine learning, memory based and neural networks. Some of the systems that ware implemented based on the neural networks approach are: POS-tagger developed by Schmid (1994) and POS-tagger developed by Antonio et al. (2001).

3. History of POS-Tagging for Non-Arabic Languages

Over the last decades there was a huge development in the POS-tagging process. The accuracy of the taggers also increased from 77% to 97%.

For English language POS-taggers were developed and spanned in the last decades, this was a result of hard work over the years. Researchers realized that POS tagging was the base for NLP (Natural language processing) systems and started to focus on developing it. NLP systems developing started in the sixties but at this time taggers were just components of NLP systems and not a separate components under the tagger name. Techniques that were used for POS-tagging process also changed from the rule-based approach in the 60's to the statistical approach in the 70's and 80's. In the 90's new approaches were used in the POS tagging process such as: machine learning, memory based and neural networks as mentioned before. Since the accuracy could not reach 100% using the statistical approach alone, the rule-based approach was combined with the stochastic approach like the work of Tapanainen and Voutilainen (1994).

Starting from 1962 Harris created a program that took English sentences as an input and produced their string-analytic elementary sentences. According to Harris (1962) sentence could be characterized by three things: constituent analysis, string analysis and transformational analysis.

Computational grammar coder (CGC) was the name of the POS-tagger that was proposed by Klein and Simmons (1963). Their tagger was one of the earliest taggers for automated POS-tagging which was developed using the

An International Journal JSCDSS E-ISSN: 2289-8603

rule-based method. They relied on small dictionaries for the words that have a unique tag and for the remaining words they used the Computational grammar coder CGC

Stolz et al. (1965) proposed the first POS-tagger that used the probabilities for determining the tag of the word. This tagger was called WISSYN grammatical coder. To calculate the probabilities a manually tagged corpus of 28,500 words was used .The tagger consisted of four phases: the dictionary, morphology, ad hoc and probability phases. The accuracy of this POS tagger reached 92.8%.

In the 1970's decade the focus was on the corpus linguistics rather than POS-tagging. With an accuracy that reached 77% an English POS-tagger by Greene and Rubin (1971) was developed using the rule-based approach. They called it TAGGIT. Brown corpus of one million words was used in this tagger. This work was considered as the most important work in the POS-tagging in that time period because it was the first tagger that provided a text corpus annotated with POS information for the tagging process.

Bahl and Mercer (1976) designed their POS-tagger. It was a probabilistic tagger. It was based on the Hidden Markov models. The accuracy of Bahl and Mercer tagger reached up to 98.6% .

In the 1980's decade, the development of the POS-taggers increased with focusing on achieving higher rates of accuracy. The POS-tagger developed by Garside (1987) tagged one million word of Lancaster-Oslo/Bergen (LOB) corpus. This tagger was known as CLAWS1 (Constituent-Likelihood Automatic Word-Tagging System). It used the hybrid technique that combined the rule-based method and the probabilistic method. HMM was used as the probabilistic component which used the lexical probabilities and contextual probabilities. For clitics, multi-word units and exception words in the rule-based approach were used.

Another POS-tagger called PARTS was developed by Church (1988). It was similar to CLAWS1 except that it used only statistical method without rule-based approach. Lexical probabilities and contextual probabilities were used in this tagger.

DeRose (1988) also developed a tagger similar to CLAWS1 and called it VOLSUNGA with an accuracy of 96%. Because of using the algorithmic tagger without using the rule of the language the accuracy was not as high as the CLAWS1, but its improvement was in the speed of the tagging process.

Inspired by the AI field and using artificial neural networks Benello et al. (1989) developed a POS-tagger used the back-propagation learning algorithm. The accuracy using this approach reached 95%. The purpose of this method was to simulate the human brain in the way it distinguishes the words. With four ambiguous words, one unambiguous word and one target word that constitute a six windows of words. By exploring the windows a complex conditional probabilities were built. Using a small context for the POS-tagging and using Brown corpus for extracting the information that constitutes the training set. The training process was on 10 sentences that ware extended to 100 sentences.

By the 1990's decade, POS-tagging process became the base for the linguistic analysis process. POS-taggers for non-English language were developed using the previous techniques and the mistakes that ware made in the POS-tagging for the English language were avoided.

Without using a tagged corpus like the previous methods Cutting (1992) developed a POS-tagger using the probabilistic method .He avoided using a tagged corpus. For English language the tagged corpus was not a problem because of the existence of the Brown and LOB corpora but for non-English languages this was an issue. The accuracy of this POS-tagger was 96%.

New POS-tagger was developed by Brill (1992) using the rule-based approach and achieved an accuracy of 97%. Brill's tagger used the rules which ware derived from the tagged corpus. This tagger belongs to hybrid method since it also used the probabilities method. Brill's tagger used an algorithm called Transformation based error driven learning (TBL) that collected the rules automatically. Using a tagged corpus for the learning process the word would be assigned to the most frequent tag by looking up the words in the corpus, if the word was not found in the corpus the tag was assigned to it according to the rules.

Kupiec (1992) developed a POS-tagger which was similar to the tagger developed by Cutting et al. In this POS-tagger there was not a training corpus so the training was performed on an untagged corpus. It achieved an accuracy of 96.36% taking into consideration all words including punctuations.

Weischedel et al. (1993) developed a POS-tagger with an improvement over the previous taggers that it required less training data. This tagger used the back propagation algorithm for the learning process. The HMM model was used, so it was considered as a probabilistic tagger

Merialdo (1994) made a comparison between the taggers with tagged corpus and the taggers with untagged corpus. He developed a tagger that used the probabilistic method. He proved that the training on the tagged corpus is better than training on untagged corpus. The tagger applied training a trigram Markov model in a similar way to HMM.

Later Using two techniques from the AI field two taggers were developed by Schmid (1994). One tagger was based on the neural networks and the other was based on the decision trees. The neural network-based tagger was known as Net-Tagger. He used a multi-layered perceptron in which the output layer consisted of all the possible tags. Assigning the tag to the word was by activating the output unit for indicating the most likely tag and deactivating other output units. The input to the neural network consisted of lexical probabilities of the tagged word and the lexical probabilities of the following words. For the training process Penn Treebank corpus of 2 million words was used. This tagger achieved an accuracy of 97.7% knowing that it allowed the ambiguous words in the output.

In the 2000's, there was not a huge progress in the development of the POS-tagging for English language. Researchers tried to improve the efficiency of the existing POS-taggers by increasing the accuracy rates. Other

An International Journal JSCDSS E-ISSN: 2289-8603

languages gained the interest of researchers by trying to develop POS-taggers for it.

Prins (2004) extended an HMM model with global contextual information; the context was incorporated separately from the POS-tags so that the number of parameters did not increase beyond practical limits. He applied his tagger on Dutch language .This method showed an increasing in the efficiency over the traditional HMM model with an accuracy reached 93.62% when tested on Alpino data.

Tamburini (2007) developed a POS-tagger as an evolution of the CORIS Tagger which was developed to tag Italian language at the University of Bologna. The proposed system composed of HMM tagger with Transformation based tagger. This tagger was used to tag CORIS/CODIS corpus successfully and it achieved an overall accuracy of 96%. In 2007 it participated to the EVALITA 2007 campaign with good results.

Experimental system architecture for POS-tagging was presented for the Italian language by Zanoli and Pianta (2009). The proposed system provided lexical and morphological information using a large tagset and it used a cascade of classifiers using support vector machine. The system achieved an accuracy of 96.06%.

Sugared (2011) developed a new algorithm that could label the data and learn from the labelled and unlabelled data and leads to more condensed model. The accuracy of the developed algorithm which was used in part of speck tagging reached 97.50%.

Table1
Summary of POS-Tagger Systems over Years for English Language

Yearr	Author/s	Methodology	Accuracy
1962	Harris	Rule based	-
1963	Klein and Simmons	Rule based	90%
1965	Stolz	Probabilities	92%
1971	Greene and Rubin	Rule based	77%
1976	Bahl and Mercer	Probabilities	98.6%
1987	Garside	Probabilities and Rule based	97%
1987	Church	Probabilities	95%-97%
1988	Derose	Probabilities	96%
1989	Benello	Neural Networks	95%
1992	Cutting	Probabilities	96%
1992	Brill	Rule based	96%
1992	Kupiec	Probabilities	96%
1993	Weischedel	Probabilities	94%
1994	Merialdo	Probabilities	N/A
1994	Schmid	Neural Networks	96%
1994	Tapanainen and Voultilainen	Probabilities and Rule based	98%
2004	Prins	Probabilities	93.62%
2007	Tamburini	Probabilities and transformation based	96%
2009	Zanoli and Pianta	Rule based	96.06%
2011	Sogaard	New Algorithm	97.50%

4. History of POS-Tagging for Arabic Language

For Arabic language different taggers had been developed by researches and companies. Companies like RDI, Sakhr and Xerox. These companies developed taggers for commercial purposes.

El-Kareh and Al-Ansary (2000) developed a hybrid tagger that used statistical method and morphological rules in the form of HMMs. Their tagger performed tests for determining the tag of the word then the user of the system could accept the current suggestion or replace it. It was called semi-automatic. El-Kareh and Al-Ansary tagger was derived from traditional Arabic grammar. It achieved an accuracy of 90%.

Shereen Khoja (2001) developed the APT system (Automatic Arabic POS-Tagger). This tagger is combined from two approaches: statistical and rule-based techniques. The APT is considered as the first tagger system for Arabic language. The tagset which was used in APT consisted of 131 tags derived from the BNC English tagset. Khoja derived her initial tagset from the grammar of Arabic language. The APT achieved an accuracy of 86 %.

Freeman (2001) used a machine learning approach and implemented Brill's POS-tagger for the Arabic language. His tagger was based on a tagged corpus. The corpus was constructed manually and it contained over 3,000 words. A tagset of 146 tags was used (Elhadj, 2009).

Maamouri and Cieri (2002) developed a POS-tagger using rule-based method. They based their Arabic tagger on automatic annotation output produced by the morphological analyzer of Tim Buckwalter. The developed tagger achieved an accuracy of 96%.

Diab et al. (2004) developed a POS-tagger for Arabic language. This tagger used the support vector machine

An International Journal JSCDSS E-ISSN: 2289-8603

(SVM) method and LDC (Linguistic Data Consortium). It consisted of 24 tagset.

Banko and Moore (2004) presented an HMM tagger for Arabic language. This tagger achieved an accuracy of 96%.

Guiassa (2006) developed a tagger that used hybrid method of rule-based and a memory-based learning method and it achieved an accuracy of 86%.

Few researchers were taken into account the structure of Arabic sentence like Shamsi and Guessoum (2006). They developed an Arabic POS-tagger for un-vocalized text with an accuracy of 97% using the HMMs.

Alqrainy (2008) developed a POS-tagger using the rule-based approach. This tagger was called AMT (Arabic Morphosyntactic Tagger). The input for AMT was untagged raw partially-vocalized Arabic corpus. The target of the tagger was to assign the correct tag to each word in the corpus producing a POS-tagged corpus without using a manually tagged or untagged dictionary. The AMT consisted of two rule components: pattern-based rules and lexical and contextual rules. The AMT system achieved an average accuracy of 91%.

Ali and Jarray (2013) used the Genetic algorithm to develop an Arabic part of speech tagging. They used a reduced tagset in their tagger.

Another POS-tagger considered the structure of Arabic sentence and combined morphological analysis with Hidden Markov Models (HMMs) developed by Elhadj (2014). The recognition rate of this tagger reached 96%.

Mohamed and Kubler (2010) developed two methods for Arabic–part of speech tagging. The two methods are: Whole word tagging and Segmentation-based tagging.

Table 2
Summary of the work that have been done in developing POS taggers for Arabic language in 2000's decade.

Year	Author/s	Methodology	Accuracy
2000	El-Kareh and Al-Ansary	Hybrid	90%
2001	Shereen Khoja	Hybrid	86%
2001	Freeman	Machine learning	-
2002	Maamouri and Cieri	Rule-Based	96%
2004	Mona Diab	Support Vector Machine	-
2004	Banko and Moore	Statistical	96%
2006	Tlili-Guiassa	Hybrid Method	86%
2006	Shamsi and Guessoum	Statistical	97%
2008	Shihadeh Alqrainy	Rule-Based	91%
2013	Bilal and Fethi	Genatic Algorithm	-
2014	Elhadj Elhadj	Statistical	96%
2014	Emad Mohamed and Sandra Kubler	Whole word tagging Segmentation-based tagging.	-

5. Conclusion

In literature, there are two main methodologies for automatic POS tagging: Rule-based methodology and stochastic (probabilistic) methodology.

Most of POS-tagging systems have been implemented using these two methodologies. Some of the existing systems combined the two methodologies to produce a hybrid one which uses the both methodologies. And some other systems used other approaches.

The above mentioned methodologies are designed for non-Arabic language. They designed for Indo-European language (e.g., English). However, these tagging systems are not fully appropriate for Arabic language. This is mainly due to the morphological and semantic differences between Arabic and other languages such as English.

In the existing literature, there are few researches have been done in developing POS-taggers for Arabic language.

Tagging with ANNs and Genetic algorithms is a new approach in Arabic natural language processing, but it had been used and applied successfully in many applications such as extracting the roots and stems for Arabic words, speech recognition, and part-of-speech prediction. These approaches need more research to apply it in Arabic language.

References

Calabrese, F. A. (2005). The early pathways: theory to practice–a continuum. Creating the Discipline of Knowledge Management, Elsevier, New York, NY, 15-20.

Capizzi, M. T., & Ferguson, R. (2005). Loyalty trends for the twenty-first century. Journal of Consumer Marketing, 22(2), 72-80.

Jakkilinki, R., Georgievski, M., & Sharda, N. (2007). Connecting destinations with an ontology-based e-tourism planner. Information and Communication Technologies in Tourism 2007, 21-32.

Aumueller, D. (2005, May). Semantic authoring and retrieval within a Wiki. In Demos and Posters of the 2nd European Semantic Web Conference (ESWC 2005), Heraklion, Greece.

Adafre, S. F. (2005, June). Part of speech tagging for Amharic using conditional random fields. In Proceedings of the ACL workshop on computational approaches to semitic languages (pp. 47-54). Association for Computational Linguistics.

An International Journal JSCDSS E-ISSN: 2289-8603

Negnevitsky, M. (2005). Artificial intelligence: a guide to intelligent systems. Pearson Education..

Alqrainy, S. (2008). A morphological-syntactical analysis approach for Arabic textual tagging.

Alqrainy, S., & Ayesh, A. (2006). Developing a tagset for automated POS tagging in Arabic. WSEAS transactions on computers, 5(11), 2787-2792.

Altunyurt, L., & Orhan, Z. (2006). PART OF SPEECH TAGGER FOR TURKISH.

Attia, M. (2006, October). An ambiguity-controlled morphological analyzer for modern standard arabic modelling finite state networks. In Challenges of Arabic for NLP/MT Conference, The British Computer Society, London, UK (Vol. 200610, No. 1.72).

Bahl, L. R., & Mercer, R. L. (1976). Part of speech assignment by a statistical decision algorithm.

Benello, J., Mackie, A. W., & Anderson, J. A. (1989). Syntactic category disambiguation with neural networks. Computer Speech & Language, 3(3), 203-217.

Brill, E. (1992, February). A simple rule-based part of speech tagger. In Proceedings of the workshop on Speech and Natural Language (pp. 112-116). Association for Computational Linguistics.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992, March). A practical part-of-speech tagger. In Proceedings of the third conference on Applied natural language processing (pp. 133-140). Association for Computational Linguistics.

Beale, M. H., Hagan, M. T., & Demuth, H. B. (1992). Neural Network Toolbox™ User's Guide. R2014a ed, 2014.

DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. Computational Linguistics, 14(1), 31-39.

Diab, M., Hacioglu, K., & Jurafsky, D. (2004, May). Automatic tagging of Arabic text: From raw text to base phrase chunks. In Proceedings of HLT-NAACL 2004: Short papers (pp. 149-152). Association for Computational Linguistics.

Attia, M. (2006, October). An ambiguity-controlled morphological analyzer for modern standard arabic modelling finite state networks. In Challenges of Arabic for NLP/MT Conference, The British Computer Society, London, UK (Vol. 200610, No. 1.72).

Church, K. W. (1988, February). A stochastic parts program and noun phrase parser for unrestricted text. In Proceedings of the second conference on Applied natural language processing (pp. 136-143). Association for Computational Linguistics.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (1992, March). A practical part-of-speech tagger. In Proceedings of the third conference on Applied natural language processing (pp. 133-140). Association for Computational Linguistics.

Beale, M. H., Hagan, M. T., & Demuth, H. B. (1992). Neural Network Toolbox™ User's Guide. R2014a ed, 2014.

Elhadj, Y. O. (2009). Statistical part-of-speech tagger for traditional Arabic texts. Journal of Computer Science, 5(11), 794.

Garside, R. (1987). The CLAWS word-tagging system.

Greene, B. B., & Rubin, G. M. (1971). Automatic grammatical tagging of English. Department of Linguistics, Brown University.

Habash, N. (2007). Arabic morphological representations for machine translation. In Arabic computational morphology (pp. 263-285). Springer Netherlands.

Habash, N., & Rambow, O. (2005, June). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 573-580). Association for Computational Linguistics.

Harris, Z. (1962). String Analysis of Language Structure. Mouton and Co., The Hague.

Jiyad, M. (2006). A Hundred and One Rules!. A short reference for Arabic syntactic, morphological & phonological rules for novice & intermediate levels of proficiency.

Jurafsky, D., & Speech, M. J. H. (2008). Language Processing. International Edition, 66-67.

Kasabov, N. K. (1996). Foundations of neural networks, fuzzy systems, and knowledge engineering. Marcel Alencar.

Kasabov, N. K. (1997). Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. Computers and Mathematics with Applications, 7(33), 136.

Khoja, S. (2003). APT: an automatic Arabic part-of-speech tagger (Doctoral dissertation, Lancaster University).

Klein, S., & Simmons, R. F. (1963). A computational approach to grammatical coding of English words. Journal of the ACM (JACM), 10(3), 334-347.

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. Computer Speech & Language, 6(3), 225-242.

McEnery, A. M., & McEnery, T. (1992). Computational linguistics: a handbook & toolbox for natural language processing. Sigma Press.

Merialdo, B. (1994). Tagging English text with a probabilistic model. Computational linguistics, 20(2), 155-171.

Al-Serhan, H. M. (2008). Extraction of Arabic word roots: An Approach Based on Computational Model and Multi-Backpropagation Neural Networks.

Van Noord, G. (2004, July). Error mining for wide-coverage grammar engineering. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (p. 446). Association for Computational Linguistics.

Schmid, H. (1994, September). Probabilistic part-of-speech tagging using decision trees. In Proceedings of the international conference on new methods in language processing (Vol. 12, pp. 44-49).

Seikaly, Z. (2007). The arabic language: The glue that binds the arab world

. Al Shamsi, F., & Guessoum, A. (2006, April). A hidden Markov model-based POS tagger for Arabic. In Proceeding of the 8th International Conference on the Statistical Analysis of Textual Data, France (pp. 31-42).

Stolz, W. S., Tannenbaum, P. H., & Carstensen, F. V. (1965). Stochastic approach to the grammatical coding of english. Communications of the ACM, 8(6), 399-405.

Tamburini, F. (2009). PoS-tagging Italian texts with CORISTagger. In Proc of EVALITA 2009. AI* IA Workshop on Evaluation of NLP and Speech Tools for Italian.

Weischedel, R., Schwartz, R., Palmucci, J., Meteer, M., & Ramshaw, L. (1993). Coping with ambiguity and unknown words through probabilistic models. Computational linguistics, 19(2), 361-382.

Zanoli, R., & Pianta, E. A multistage PoS-tagger at the EVALITA 2009 PoS-tagging Task.

Søgaard, A. (2011, June). Semisupervised condensed nearest neighbor for part-of-speech tagging. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2 (pp. 48-52). Association for Computational Linguistics.

Ali, B. B., & Jarray, F. (2013). Genetic approach for arabic part of speech tagging. arXiv preprint arXiv:1307.3489.

Mohamed, E., & Kübler, S. (2010, June). Is Arabic part of speech tagging feasible without word segmentation?. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational

An International Journal JSCDSS E-ISSN: 2289-8603

Linguistics (pp. 705-708). Association for Computational Linguistics.
Schmid, H. (1994, August). Part-of-speech tagging with neural networks. In Proceedings of the 15th conference on Computational linguistics-Volume 1 (pp. 172-176). Association for Computational Linguistics.
Prins, R. (2004, July). Beyond N in N-gram Tagging. In Proceedings of the ACL 2004 workshop on Student research (p. 61). Association for Computational Linguistics.
Tamburini, F. (2009). PoS-tagging Italian texts with CORISTagger. In Proc of EVALITA 2009. AI* IA Workshop on Evaluation of NLP and Speech Tools for Italian.
Elhadj, Y. O., Abdelali, A., Bouziane, R., & Ammar, A. H. (2014, November). Revisiting Arabic Part of Speech Tagsets. In Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on (pp. 793-802). IEEE.
Abbas, Q. (2014). Semi-semantic part of speech annotation and evaluation. LAW VIII, 75.
Schneider, G., & Volk, M. (1998). Adding manual constraints and lexical look-up to a Brill-tagger for German. In Proceedings of the ESSLLI-98 Workshop on Recent Advances in Corpus Annotation, Saarbrücken.
Perez-Ortiz, J. A., & Forcada, M. L. (2001). Part-of-speech tagging with recurrent neural networks. Universitat d'Alacant, Spain.
Chanod, J. P., & Tapanainen, P. (1995, March). Tagging French: comparing a statistical and a constraint-based method. In Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics (pp. 149-156). Morgan Kaufmann Publishers Inc..
Mohamed, E., & Kübler, S. (2010, May). Arabic Part of Speech Tagging. In LREC.
Ku, H., & Francis, W. N. (1967). Computational Analysis of Present-Day {A} merican {E} nglish.
MacWhinney, B., & Snow, C. (1985). The child language data exchange system. Journal of child language, 12(02), 271-295.
Aarts, J., van Halteren, H., & Oostdijk, N. (1998). The linguistic annotation of corpora: The TOSCA analysis system. International journal of corpus linguistics, 3(2), 189-210.
Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 19(2), 313-330.
Al-Sulaiti, L., & Atwell, E. S. (2006). The design of a corpus of contemporary Arabic. International Journal of Corpus Linguistics, 11(2), 135-171.
Das, B. R., Sahoo, S., Panda, C. S., & Patnaik, S. (2015). Part of Speech Tagging in Odia Using Support Vector Machine. Procedia Computer Science, 48, 507-512.
Stenström, A. B., Andersen, G., & Hasund, I. K. (2002). Trends in teenage talk: Corpus compilation, analysis and findings (Vol. 8). John Benjamins Publishing.
Calciu, R. H. Semantic change in the age of corpus linguistics. EDITORIAL SECRETARY, 45..
Maamouri, M., & Bies, A. (2004, August). Developing an Arabic treebank: Methods, guidelines, procedures, and tools. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based languages (pp. 2-9). Association for Computational Linguistics.
Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn treebank: an overview. In Treebanks (pp. 5-22). Springer Netherlands.
Maamouri, M., Bies, A., Buckwalter, T., & Mekki, W. (2004, September). The penn arabic treebank: Building a large-scale annotated arabic corpus. In NEMLAR conference on Arabic language resources and tools (Vol. 27, pp. 466-467).

An International Journal JSCDSS E-ISSN: 2289-8603