Volume 2, No. 1, Jan-Feb 2011
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Web-Based Arabic Language Orthography Enrichment System

Chuttur Y Mohammad
Indiana University
Bloominton, Indiana, USA
mchuttur@indiana.edu

Abstract: Arabic language presents several difficulties to new learners. In this paper, a novel approach consisting of a web based system that can automatically present orthographic details of Arabic text in order to assist Arabic language learners is presented. The different components of the proposed system are described, and results obtained by testing the system with short Arabic stories show that the system can effectively process Arabic text to provide useful orthographic information. Future enhancements to the system are also discussed.

Keywords: part-of-speech tagging, machine learning, Arabic orthography, online learning.

I. INTRODUCTION

Arabic is a morphologically complex language that is notoriously difficult to read due to two main difficulties: (1) Arabic is written without the short vowels, and this renders the text ambiguous, for example, when a learner encounters the word رصم (mSr[1]) she will most probably not be able to tell whether it is the proper noun miSor (Egypt), the adjective muSir (insistent), or the verb maS~ra (to Egyptianize), among many more possible vocalizations of the word, (2) the Arabic orthography stacks many prefixes at the beginning and end of the word, and this makes it difficult for learners to know which word they are dealing with. For example, the word (ولمعلاب) wbAlEml, which means "And with the work", comprises a conjunction (w), a preposition (b), a definite article (Al), and a noun (Eml). This makes it even harder for the learner to find the word in a proper dictionary. It is possible that by enriching the text with basic information such as the different parts of the word (stems, affix), and the function of each part (e.g. conjunction, feminine singular marker), the reading process can be facilitated. This task has been achieved by building a web-based tool that reads Arabic text and automatically annotates every word of the text, yielding a highly enriched output text that can aid a non-native Arabic speaker to read and learn the language efficiently. To situate the context of this work, some relevant work carried out in this area are discussed followed by a presentation of the architecture of the developed system, which comprises different modules such as a word segmenter, a Part of Speech (POS) Tagger, and a visual renderer output module. Brief explanations for each module are also given, and after providing a snapshot of the output of the system, further research directions that are presently being worked on are provided.

II. RELATED WORK

Although Arabic word segmentation, and Part of speech tagging have been the subject of some research; to date, there exists no current system that combines both techniques for assisting Arabic language learners. Previous works in Arabic word segmentation, and Part of speech tagging have remained distinct in their disciplines as described here. Darwish [1], for example, used a corpus of stems and a list of prefixes and suffixes to strip affixes from word beginnings and ends. Darwish assumed that a word would carry at most one suffix and one prefix, and to operationalize his idea, he grouped combinations of affixes together. For example, the word "wbAlEml" above is not treated as w+b+Al+Eml , but as wbAl+Eml. Darwish reports and accuracy of 92.7% on a 9606 word corpus. Diab et al [4], and Habash and Rambow [5] instead used a Support Vector Machine (SVM) approach, and both reported accuracies of over 99%.

For Arabic Part of Speech Tagging, all studies assumed that the input to the tagger is perfectly tokenized, and thus used the gold standard tokenization distributed in the Arabic Treebank (ATB) [6]. Diab et al used the Reduced tag Set, which is a minimal tag set distributed with the ATB, and reported an accuracy of 95.5 on all tokens drawn from the ATB. Habash and Rambow [5] used a morphological analyzer to aid their SVM-based POS tagging, and they report an accuracy of 97.6 on gold standard tokenization using a Reduced Tag Set. Van den Bosch et al [7] used Memory-based learning for both morphological analysis and POS tagging of Arabic. But unlike Habash and Rambow [5] and Diab et al [4], they use mostly whole words in their approach, as they keep the words in the way they looked in the ATB without re-attaching the conjunctional prefixes and the pronominal and prepositional suffixes. Van den Bosch et al. report a total accuracy of 91.5% with 93.3% accuracy on known words and 66.4% accuracy on unknown words.

III. ARABIC ORTHOGRAPHY ENRICHMENT SYSTEM

Figure 1 gives the general architecture of the proposed Arabic language orthographic enrichment system. The system that combines both Arabic word segmentation and Part of speech tagging and thus, consists of several modules: a word segmenter, a part of speech tagger, an annotation module, and a visual render module that work in sequence to finally produce a web page with enriched orthographic annotations

[1] Throughout this abstract, Arabic words are presented in Buckwalter transliteration.

from raw Arabic text as input. Furthermore for each Arabic word selected, the reader is provided with a hyperlink to Google translate (http://translate.google.com/) that automatically gives the English translation of the word. The reader may also select among other languages to translate the Arabic word on the Google translate website. Each module was developed and tested individually as described in the following sections.

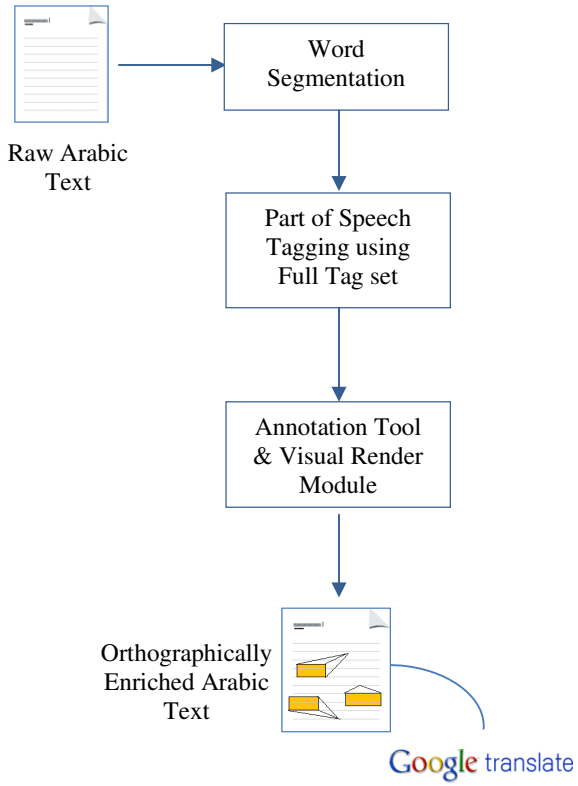

Figure 1. General architecture for the system

A. *Word Segmentation Module*

The word segmenter for Arabic was built using the memory-based approach and the TiMBL implementation [2]. The word segmenter takes as input an Arabic text, and returns the text with word boundaries marked. For example, given the sentence:

وسوف يقوم سيادة الرئيس بافتتاح عدد من المشروعات المهمة غدا ويرافقه وزير الإسكان

the segmenter returns:

و+سوف ي+قوم سياد+ة ال+رئيس ب+افتتاح عدد من ال+مشروع+ات ال+مهم+ة غد+ا و+ي+رافق+ه وزير ال+إسكان

The segmenter scores an accuracy of 98.1% across 5 fold cross validation.

B. *Part of Speech Tagging Module*

Since there is no POS tagger available that provides a reasonably detailed analysis of the words, the tagger is built using MBT, the memory-based tagger [3]. The tagger takes as input the output of the word segmenter, and produces the input enhanced with grammatical categories. The tagger was on the Arabic Full Tag Set, but modified the tags to be more expressive for a language learner. For example, a word like wbm$rwEAthm (وبمشروعاتهم) is tagged as in Table I.

Table I. More Informative Tags for the Word وبمشروعاتهم

Word Part	*Tag*
و	conjunction possibly meaning and
ب	preposition, possibly meaning with
مشروع	noun
ات	Noun suffix indicating a feminine plural
هم	Possessive pronoun for the masculine plural = their

The part of speech tagging module was experimented with tagging using whole words and tagging using segmentation as a preprocessing step. It was found out that performing segmentation before POS tagging yields better results when the number of previously unknown words is large, while tagging on whole words yields better results when the number of previously unknown words is small.

Table II presents the Part of Speech Tagging experiments in a 5 fold cross validation setting over two parts of the Arabic Treebank (P1V3 and P3V1). The table presents evaluations in terms of the segment accuracy rate (SAR) and word accuracy rate (WAR).

Table II. POS Tagging Accuracy across 5 Folds

	Gold Standard segmentation		Developed segmentation		Whole Words
	SAR	WAR	SAR	WAR	WAR
Avg.	96.718%	94.910%	94.598 %	93.406%	93.933%

Table III, on the other hand, shows the results obtained on unknown words alone. Therefore, when performing POS tagging for a language genre suitable for learners, it is suggested that segmentation be performed first, as this belongs to a different genre than the Arabic Treebank, implying that the number of unknown words may be large.

Table III. POS Tagging Accuracy across 5 Folds

Gold Standard Segments	Using developed segmenter	Using Whole Words
84.25%	70.50%	65.50

C. Annotation and Visual Renderer Module

The Arabic orthographic enrichment system works through the pipeline illustrated in Figure 1 above to provide the learner with information regarding the structure of the Arabic word. The visual renderer engine essentially uses the output from the POS module to provide a java-script enabled web page that lets the learner select any word from a text written in Arabic, and obtain an annotated window, similar to the example shown in Table 1, with the units of the word, its grammatical structure, and a link to Google Translator that gets the word translation instantly. The window is also color-coded with stems presented in a different color than the affixes, in order to make it easier for the learner to capture the essence of the word. Figure 2 shows a snapshot of the enriched annotated text obtained when a user selects multiple words from a short story written in Arabic. An example of the output from the system can be seen at http://jones.ling.indiana.edu/~emadnawfal/augmentedDuck.html. (Character Encoding need to be set to UTF-8).

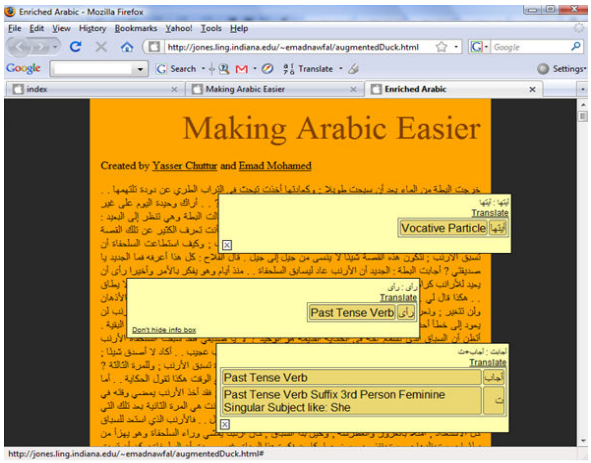

Figure 2. Output showing a short story written in Arabic and additional orthographic information selected by a reader.

IV. System Testing and Evaluation

Eight participants who enrolled in an introductory course in Arabic language were chosen to participate in using and evaluating the system. Although, evaluation was centered on whether learners have improved understanding of Arabic using the system, a short questionnaire also requested participants to evaluate the usability of the system. All participants were between 18 and 25 years old and the gender was split between males and females equally. Participants were recruited on a first come first serve basis following announcement for participation in the evaluation through email invitation to students registered in an introductory Arabic language course. Since the system requires a minimal understanding of Arabic, the experiment was carried out at the end of the introductory Arabic language course that participants were enrolled into. All participants were native English language speakers and although not fluent in Arabic, they all had sufficient experience reading and understanding Arabic language suitable for testing the system. Overall GPAs for each participant by the end of the introductory course showed that they all had average knowledge of the Arabic language.

The task consisted in reading a short story written in Arabic and to then answer six questions that pertained to the text. These questions tested the comprehensibility of the text and therefore served as evidence regarding whether the system is indeed assisting learners in better understanding a foreign language like Arabic.

Participants were assigned to two groups such that in one group of four participants (2 males and 2 females) were supplied with the story written in Arabic on a printed document, while the second group was able to read the same story online using the system developed for Arabic orthographic enrichment. At the end of the experiment, participants in group 2 were asked to rate the usability of the system on a scale of 1 to 5 with 1 being very unsatisfied and 5 being very satisfied. Open ended questions asking for any difficulty using the system was also requested from participants in group 2.

V. Results and Discussion

The experiment lasted for roughly an hour and all results obtained were carefully analyzed by a native Arabic speaker. Overall, the mean scores for group 2, who used the system, obtained higher scores that those participants who had only a paper print version of the story. However, it was noticed that on average, participants using the system spent more time in answering all the questions compared to those who were not exposed to the system. The time noted was on average 38.2 minutes for group 1 and 50.5 minutes for group 2. It was suspected that participants using the system spent too much time reading through unfamiliar terms and seeking help from Google translate.

Regarding the usability of the system, all participants rated the system as highly satisfactory, but one participant noted that sometimes, the different annotations showing the orthographic description of a word will hide the actual text being read and this often made her lose track of where she were in the actual text. As a recommendation, she suggested that all annotations should be made available in a separate frame within the same window that carries the Arabic text.

The results obtained show promising application for the system for Arabic learners, However, given the small sample of participants (n=8), the results obtained cannot be generalized and require further evaluation with more participants. But the results obtained are sufficient to show the usability of such a system for assisting new learners of Arabic in comprehension and learning the language. It was also interesting to note that participants spent more time using the system to complete the experiment. As one of the participant suggested, this could be due to the reader switching back and forth between different windows such that time is spent relocating the point at which they were in the original text. However, the added benefit as noted was that participants performed much better using the system than if they read the same text without any orthographic enrichment.

VI. Conclusion and Future Works

This paper has presented a web based system that can efficiently process text written in Arabic by adding rich orthographic details to the text in order to assist readers or learners of Arabic. The proposed system combines two techniques namely, Arabic word segmentation and Part of speech tagging that have not been combined and tested previously. By applying both techniques in the context of

Arabic language learning and comprehensibility, it was found that participants performed better.

Currently the system is made of separate components that require an advanced user, familiar with UNIX platform and Python to be able to upload their own texts written in Arabic and run each of the modules shown in figure 1 in a pipeline. In the future, the intent is to deploy the system to a larger number of users with minimal computing experience. Further development planned for the system therefore is considering the possibility of combing all the separate modules of the system in one application such that a user can easily enter her own Arabic text, or a url pointing to an Arabic text, and output an enriched version of the input text.

Further research is also on the way to apply the system to a wide variety of texts covering various topics written in Arabic and to evaluate the effect of using the enriched text on three main tasks (1) Ease of reading (2) Text comprehension and, (3) Text translation. The system will be tested on intermediate learners of Arabic for these tasks and eventually the experiment will also be expanded to include beginners with simpler Arabic text. Other enhancements planned for the system include integrating vocalization and target color-coding in the system.

VII. ACKNOWLEDGMENT

The author wishes to express his gratitude and sincere thanks to Dr Emad Mohamed Nawfal, Suez University, Egypt, for his useful input for this project.

VIII. REFERENCES

[1] Darwish, Kareem. 2002. Building a shallow Arabic morphological analyser in one day. In ACL02 Workshop on Computational Approaches to Semitic Languages, Philadelpia, PA. Association for Computational Linguistics.

[2] Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. 2007a. TiMBL: Tilburg Memory Based Learner, version 6.1, Reference Guide. ILK Research Group Technical Report Series no. 07-07.

[3] Daelemans, W., Zavrel, J., Van den Bosch, A., and Van der Sloot, K. 2007b. MBT: Memory-Based Tagger, version 3.1, Reference Guide. ILK Technical Report Series 07-08.

[4] Diab, Mona ; Hacioglu, Kadri ; and Jurafsky, Daniel. 2004. Automatic tagging of Arabic text: From raw text to base phrase chunks. In 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04), Boston, MA.

[5] Habash, Nizar and Rambow, Owen. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. Proceedings of the 43rd Annual Meeting of the ACL, pages 573–580, Ann Arbor, June 2005. Association for Computational Linguistics

[6] Maamouri, Mohamed; Bies, Ann; and Buckwalter, Tim. 2004. The Penn Arabic Treebank: Building a large scale annotated Arabic corpus. In NEMLAR Conference on Arabic

Language Resources and Tools, Cairo, Egypt.

[7] Van den Bosch, Antal; Marsi, Erwin; and Soudi, Abdelhadi. 2007. Memory-based Morphological Analysis and Part-of-speech Tagging of Arabic in Soudi, A.;van den Bosch, A. and Neumann, G. (eds.). Arabic Computational Morphology, Springer.