An Evaluation of AMIRA for Named Entity Recognition in Arabic Medical Texts

Saad Alanazi[1, 2], Bernadette Sharp[2] and Clare Stanier[2]

[1] College of Computer Science and Information, Aljouf University, Skaka, Saudi Arabia
saad.alanazi@research.staffs.ac.uk

[2] Faculty of Computing, Engineering and Technology, Staffordshire University, Beaconside, Stafford ST18 0AD, UK
{B.Sharp, C.Stanier} @staffs.ac.uk

Abstract. A study is carried out to evaluate the AMIRA tool which has been used widely to pre-process Arabic texts for natural language processing tasks. AMIRA is used in our study to tokenise and POS tag our Modern Standard Arabic medical texts. AMIRA includes a tokeniser, POS tagger, and a base phrase chunker. The AMIRA tokeniser has achieved 91.22%, 87.15% and 89.13% for precision, recall and F-measure, respectively, while AMIRA POS tagger achieved 84.09% accuracy. The most common errors in the tokeniser outputs were in the words where the first letter after the ال (Al) determiner is ل (L). With respect to the POS tagging, AMIRA underperformed in the following categories: broken plurals, adverbs, adjectives and genitive nouns.

1 Introduction

The term "Named Entity", which was coined for the Sixth Message Understanding Conference (Grishman & Sundheim 1996) was initially applied to information extraction tasks aimed at extracting names of person, organisation and locations as well as numeric and percent (e.g. time, date, money) expressions from structured and unstructured documents. This task was not only recognised as essential step of information extraction but became a focus of study for many researchers.

This paper focuses on text tokenisation and part-of-speech tagging (POS), two crucial steps in many natural language processing applications and, in particular, in named entity recognition. The first task is tokenisation which aims to convert text into tokens, where tokens are one or more characters that express an independent linguistic meaning, and roughly correspond to words. The tokenisation task is crucial because errors made in this phase can propagate into later phases and lead to serious problems. It may seem less challenging in the context of some languages, such as English, where a single space or punctuation is used to split sentences into words (tokens). However, it is very challenging in some languages, like Chinese, Japanese, and Thai, which do not use spaces to split sentences into words (Peng et al., 2004). It is a challenging and non-trivial task in the Arabic language as word tokens cannot be delimited solely by a blank space because Arabic words are often ambiguous in their morphological structure. The

aim of the second task is POS tagging which assigns an appropriate POS tag to every token in the input data (Voutilainen, 2003). As Arabic has a very rich and complex morphology a word can carry not only inflections but also clitics, such as pronouns, conjunctions, and prepositions. A single stem may correspond to thousands of different word forms (Habash, 2010; Mohamed & Kübler, 2010).

The aim of our research is to extract information about symptoms, treatment and drugs relevant to cancer from Arabic medical literature. We have used the AMIRA tool developed at Stanford University (Diab, 2009) in our tokenisation and POS tasks. This paper discusses the problems and issues encountered in applying AMIRA. Section 2 explains the challenges related to tokenisation and POS of Arabic texts. Section 3 reviews previous work and section 4 describes the data set, the experimental set up and discusses the results. Section 5 presents our final findings.

2 Challenges of Arabic Language Processing

Arabic has many traits which, make building an effective tokenising and POS tagging tool a very challenging task. Some of these main challenges are described below.

2.1 Agglutination

The Arabic language has an agglutinative nature and this results in different patterns, which can create many lexical variations. It has a very systematic, but complicated morphology. This is seen with words that comprise prefixes, a stem or a root, and sometimes even more than one, as well as suffixes with different combinations. There are also clitics, which in most languages, including English, are treated as separate words; however in the Arabic language, they are agglutinated to words (Farghaly and Shaalan, 2009). For instance, a phrase in English, such as "and they will write it" can be split into five tokens, while in Arabic this is expressed in one word وسيكتبونها (wsyktbonha). As this example demonstrates, the conjunction “and” and the future marker “will” are represented as prefixes by the letter و and س, respectively, while the pronouns “they” and “it” are represented by the suffixes ون and ها, respectively. Because of the complex morphological structure of the Arabic language, the tokenisation process is a difficult and challenging task.

2.2 Short Vowel Absence

Diacritics can be found in Arabic text, which is a representation of most vowels that affect phonetic representation. This lends an alternative meaning to the same word. Consequently, disambiguation in the Arabic language is a difficult task because it is may be written without diacritics (Alkharashi, 2009). For instance, the word كتب without using diacritics could mean the noun “books” or the verb “to write”; therefore, determining the appropriate POS tag is difficult in the absence of diacritics.

2.3 Rich Morphology

Arabic has a very rich morphology. As a result, a vast number of words can be derived from only one root. For instance, the following words have been derived from the root ك ت ب (k t b): كتب (wrote), كتاب (book), كاتب (writer), كتبة (writers – broken plurals), كُتّاب (writers – broken plurals), مكتب (office),مكتبات (offices), مكتبة (bookstore), مكتوب (written), كُتيب (booklet), كاتبون (writers- masculine), كاتبات (writers- feminine), كتيبة (Battalion), and so on. Consequently, the tag set can potentially be huge and can reach over 330,000 tags for untokenised words (Habash, 2010), an additional challenge for Arabic POS tagging.

3 Previous Research

The tokenisation process is often discussed as a part of several existing morphological analysers, such as the Buckwalter Arabic morphological analyser (BAMA), AMIRA (Diab, 2009), MADA+TOKAN, Khoja stemmer and the tri-literal root extraction algorithm (Al-Shalabi et al, 2003). BAMA uses pre-stored dictionaries of words, stem and affixes constructed manually, as well as truth tables to determine their correct combinations (Buckwalter, 2004; (Buckwalter, 2002). BAMA consists of three parts: lexicon, compatibility tables, and an analysis engine. All the prefixes, suffixes, and stems are gathered in a different lexicon. The task of the compatibility table is to determine whether the morphological units (prefix-stem- suffix) are permitted to occur all together or not. The analysis engine produces different morphological analyses such as POS tag, lemma, and morpheme analyses. AMIRA and MADA, both use a support vector machine (SVM) to perform the tokenisation of Arabic words. The AMIRA tool (Diab, 2009) which was developed at Stanford University, includes a tokeniser, POS tagger, and a base phrase chunker. AMIRA uses a fixed size window of +/- five letters; all letters tags within the window are used as features to feed the SVM algorithm. AMIRA provides the user with a choice of three tagging schemes: Bies, ERTS, and ERTS_PER tag sets. In the MADA+TOKAN system MADA which is the morphological analyser makes use of orthogonal features and a list of potential analyses provided by BAMA to select the most appropriate analysis of each word. TOKAN uses morphological generation to recreate the word after splitting off its clitics (Habash et al., 2009). In the Khoja stemmer (Khoja, 1999), the longest prefix and suffix are removed from the word, and then the remainder of the word is matched with the patterns of different nouns and verbs. The stemmer makes use of a list of all diacritic characters, punctuation characters, definite articles, and stop words (Larkey & Connell 2001). Al-Shalabi et al. (2003) have developed a tri-literal roots extraction algorithm that does not depend on any pre-stored information, but assigns mathematical weight to the position of the letters in a word. Higher weights are assigned to the letters at the beginning and at the end of the word and lower weights to root letters.

A comparative analysis of the three stemmers, Khoja stemmer, BAMA, and tri-literal root extraction algorithm, was carried out by Sawalha and Atwell (2008). These three systems were applied to two distinct documents: a newspaper and a chapter

from the Qur'an, each containing about 1000 words. The three stemming algorithms have generated correct analysis for simple roots that do not require detailed analysis. The performance is computed using a majority voting procedure in selecting the most common root among the list of words and their roots. Their analysis showed about 62% average accuracy rate for Qur'an text and about 70% average accuracy for newspaper text.

4 Experimental Study with AMIRA

The accuracy of the stemmers may not be an important issue for information retrieval systems but it is vital for named entity recognition applications. Our approach to extracting specific named entity from cancer documents consists of four main stages: pre-processing, data analysis, feature extraction, and classification stages. The pre-processing stage (in dashed line) covers the data tokenisation and POS tagging approach, which is the focus of this paper. The resulting tokens and their grammatical tags are transferred into a set of features which are then used as inputs for the classification phase. It is proposed to use Bayesian Belief Network to train and classify the extracted features which will then become the recognised entities. Any errors encountered in the early processing of texts have to be rectified to avoid their propagation in subsequent tasks and to produce a reliable training system. Figure 1 displays our named entity recognition system architecture.

In order to perform the text tokenisation task, the AMIRA tool was used as it accepts raw Arabic texts as input and allows the user to choose between different tokenisation schemes.

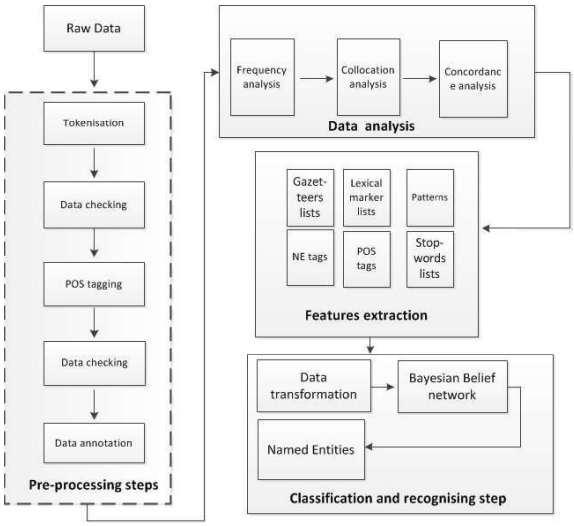

Fig. 1. The NER system architecture

4.1 Data Description

The data for our study is based on Modern Standard Arabic texts extracted from the King Abdullah Bin Abdulaziz Arabic Health Encyclopaedia (KAAHE) website. KAAHE was initiated through the collaboration between the King Saud Bin Abdulaziz University for Health Sciences and the Saudi Association for Health Informatics and further developed by the National Guard Health Affairs the Health on the Net Foundation and the World Health Organisation. KAAHE became the official health encyclopedia in May 2012 (Saudi E-health Organisation, 2012). KAAHE is a reliable health information source, contains abundant information written in an easily understandable language appropriate for users from various community groups (Alsughayr, 2013).

4.2 Tokenisation Task

AMIRA was applied to 26 articles with a total of 5119 tokens. Each article is related to a specific type of cancer. AMIRA allows the user to determine the tokenisation scheme from the different existing schemes. Different prefixes such as conjunctions, future markers and prepositions are selected to be split into parts. The Al determiners and suffixes are not tokenised because this increases the ambiguity and sparsity of the text, as there are more than 127 suffixes in Arabic (Sawalha and Atwell, 2009). Figure 2 displays a sample of the tokenisation result where errors are highlighted in grey.

Uterine cancer is cancer that begins in the uterus. This program will focus on the most common type of uterine cancer, which is endometrial cancer. Endometrial cancer begins in the lining of the uterus. Cancerous cells spread to different parts of the body through blood vessels and lymph channels. It is usually impossible to specify the .cause of cancer in an individual patient	سرطان الرحم هو السرطان الذي يبدأ في الرحم. يهتم هذا البرنامج بالنوع الأكثر شيوعا من سرطان الرحم، وهو السرطانة البطانية الرحمية. تبدأ السرطانة البطانية الرحمية في بطانة الرحم الداخلية. تنتشر الخلايا السرطانية إلى أجزاء مختلفة من الجسم عن طريق الأوعية الدموية و القنوات الالمفية. و يكون من المستحيل تحديد السبب الدقيق ل الإصابة بالسرطان لدى مريض ب عينه عادة

Fig. 2. A sample of the tokenization task result

In the above example, AMIRA missed tokenising the words: بالنوع (*bAlnwE* - type) and بالسرطان (*bAlsrTAn* – by cancer) which starts with the preposition ب (b) and the word وهو (*whw* – and it) which starts with the conjunction و (w). On the other hand, AMIRA tokenised the word اللمفية (*Allmfyp* -lymphatic),which does not need to be tokenised, by adding ا (A) letter after the determiner ال (Al) so the wrong result of tokenisting this word is الالمفية (*AlAlmfyp*).

We evaluated the results of AMIRA's tokenization result in terms of three measures, precision, recall and F-measure using the following equations:

$$Precision = \frac{the\ number\ of\ words\ which\ have\ been\ tokenised\ correctly}{the\ number\ of\ words\ which\ have\ tokenised}$$

$$Recall = \frac{the\ number\ of\ words\ which\ have\ been\ tokenised\ correctly}{the\ number\ of\ words\ which\ need\ to\ be\ tokenised}$$

$$F - measure = \frac{2 * recall * precision}{recall + precision}$$

The AMIRA tool has achieved 91.22%, 87.15% and 89.13% for precision, recall and F-measure, respectively. Two categories of errors are identified:

- False positive errors that occur when AMIRA tokenises a word that does not need to be tokenised.
- False negative errors that occur when AMIRA misses tokenising word that needs to be tokenised.

One of the most common false positive errors was tokenising words where the first letter after the ال (Al) determiner is ل (L). Examples of these words are: اللعابية (*AllEAbyp* - salivary), اللمفية (*Allmfyp* - lymphatic), اللوزتين (*Allwztyn* - tonsils) and اللوكيميا (*AllwkymyA* - leukemia). Some of these errors may be related to the limited data set used by AMIRA's classifier. These errors were corrected manually before moving to the next task. AMIRA adds a ا (A) letter after the determiner in these words so the wrong results of tokenising these words are الالعابية (*AlAlEAbyp*), الالمفية (*AlAlEAbyp*), الالوزتين (*AlAlwztyn*), and الالوكيميا (*AlAlwkymyA*). A proposed solution for this error is not to tokenise any words that have a double letter ل (L), unless the double ل (L) is the first two letters, or to insert a good number of examples of these words into the training data if the tokenisation system is using a machine learning technique, as with AMIRA. Regarding false negative errors, the main words were those that started with the ب (b) preposition. Examples of these words are: بالسرطان (*bAlsrTAn* -by cancer), بحسب (*bHsb* - according to), بالدهون (*bAldhwn* - with fats), باليود (*bAlywd* - with iodine). It is possible to split the ب (b) preposition if the following letters are the determiner ال (Al). This is because Arabic words which start with بالـ (bAl), where the ب (b) is an original letter of the word, are very uncommon. In order to examine how common these words are, the ANERcorp corpus, which consists of around 150,000 tokens (Benajiba et al., 2007) was used. Among the ANERcorp, 1104 words start with بالـ (bAl). However, in only 21 of these is بالـ (bAl) part of the original word, and nine words of the 21 words are actually non-Arabic. The rest of the words are a repetition of only four Arabic words which are بالغة (*bAlghp* - exaggerate), بالغ (*bAlgh* - adult), بال (*bAl* - shabby) and بالي (*bAly* - shabby). Creating a gazetteer for words which start with بالـ (bAl) when the ب (b) is an original part of the word, would assist the tokenisation of such words.

4.3 POS Tagging

AMIRA is also applied to perform POS tagging. Three different tag sets are available: Bies tag set, Extended Reduced tag set (ERTS) and Extended Reduced tag set + person information (ERTS_PER). The Bies tag set was developed by Ann Bies and Dan Bikel and consists of 24 tags. It ignores certain Arabic distinctions, for example, it treats the dual form, a common form in Arabic language, as a plural. It also can not specify gender in both verbs and nouns. The ERTS tag set has 72 tags and provides additional morphological features to the Bies tag set, and can handle number (singular/dual/plural), gender (feminine/masculine) and definiteness (the existence of the definite article or not). In addition to the tags in the ERTS tag set, the

Fig. 3. A sample of the POS tagging task result

ERTS_PER specifies the use of the first, second and third person voice. The ERTS, which was selected for the POS tagging task, has many relevant morphological features to our corpus while Person information is a less important feature as our data only has the third person voice. Figure 3 displays a sample of the POS tagging task result.

In the above example, AMIRA assigned a noun tag to the place adverbs خلف (behind) and أمام (in front of). It also assigned an adjective tag to the genitive noun المعدة (stomach). Amira also failed in assigning a plural noun tag NNS to the word عوامل (factors). We evaluated the results of AMIRA's POS tagging in terms of the accuracy. POS tagger accuracy is the number of correctly tagged tokens divided by the total number of tokens. AMIRA achieved an accuracy of 84.09%. However, Arabic POS taggers still need more research efforts to improve the accuracy and reach a standard equal to Stanford POS tagger for English language which has achieved 97.3% accuracy (Manning, 2011). The areas where AMIRA performed less than the average is explained below.

- **Broken plurals**

Arabic has three types of plurals: the broken plural, the sound masculine plural and the sound feminine plural. The most used type is the broken (irregular) plural, constituting about half of all plurals in Arabic (Habash, 2010). AMIRA has limited capability to assign an appropriate POS tag to broken plurals, as 32.02% of AMIRA errors are related to broken plural words. For instance, AMIRA assigns a singular feminine word tag (DET_NN_FS) to the broken plural words الأوعية (utensils), الأنسجة (tissues) and الأقنية (ducts). It also failed to assign a plural noun tag (NNS) to most of the other broken plural words. Examples of these words are الأطباء (doctors), سُبُل (ways) and خلايا (cells). Broken plurals can be formed using more than 20

morphological patterns. Furthermore, an Arabic word might have more than one plural. For instance, the word أسد (lion) has five different broken plural forms (آساد - أسود - أسُد – أسَدة – أُسْد). Therefore, it can be quite difficult to identify a solution for broken plural POS tagging. We propose to improve the performance of broken plurals POS tagging by using machine learning classifier techniques such as neural networks, or decision tree. In the literature, Goweder et al (2004) examine different methods in order to identify the broken plural. Then concluded that the dictionary and decision tree methods achieved the highest results in identifying broken plurals.

- **Adverbs**

In Arabic, there are two main types of adverb: those describing time and others referring to place or location. AMIRA assigned a noun tag (NN) to most adverbs in our corpus. Examples of these adverbs are: خلف (behind), أسفل (at the bottom of) and بعد (after). We propose to create an adverb gazetteer and use it as a binary feature to feed the machine learning classifier.

- **Adjective and genitive nouns**

One of the most frequent errors in AMIRA's POS output is assigning an adjective tag (JJ) to genitive nouns (المضاف إليه). For instance, AMIRA assigns a JJ tag to the word 'stomach' in the phrase سرطان المعدة (cancer of the stomach), the word 'patient' in the phrase فرصة المريضة (the patient's chance) and the word 'appetite' in the phrase نقصان الشهية (loss of appetite). There are some grammatical differences between adjectives and genitive nouns, in Arabic grammar. Adjectives and the nouns that they modify must agree in number (singular/dual/plural), mood (indicative/subjunctive/ genitive) and in indefiniteness and definiteness (presence of the definite article). In the above examples, the adjectives and the nouns that they modify disagree in both mood and the indefiniteness and definiteness. Using these grammatical differences as features in the data training phase will improve the task of differentiation between adjectives and genitive nouns.

5 Conclusion

Tokenisation and POS tagging are two important tasks used at early stages of named entity recognition systems. Whilst these tasks may be seem less challenging when processing English texts, many challenges face their implementation for Arabic texts because of the complex morphological structure of the Arabic language. This paper has described some of these challenges encountered by the use of AMIRA to tokenise and POS tag articles related to cancer extracted from the health encyclopedia. The AMIRA tokeniser has achieved 91.22%, 87.15% and 89.13% for precision, recall and F-measure, respectively, while AMIRA POS tagger achieved 84.09% accuracy. The most common errors in the tokeniser output were in the words where the first letter after the ال (Al) determiner is ل (L). With respect to the POS tagging, the areas where AMIRA underperformed include broken plurals, adverbs,

adjectives and genitive nouns. Some of these errors can be addressed using machine learning techniques which will be the subject for future work.

Acknowledgment

This research is supported by Aljouf University, Saudi Arabia and Staffordshire University, UK.

References

Alkharashi, I. (2009) Person named entity generation and recognition for Arabic language. In: *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, pp.205–208.

Al-Shalabi, R., Kanaan, G., & Al-Serhan, H. (2003*). New approach for extracting Arabic roots*. Paper presented at the International Arab Conference on Information Technology (ACIT'2003), Alexandria, Egypt.

Alsughayr A. (2013) King Abdullah Bin Abdulaziz Arabic health encyclopedia (www.kaahe.org): A reliable source for health information in Arabic in the internet. Saudi J Med Med Sci; 1: 53-4

Benajiba, Y., & Paolo, R. (2007) ANERsys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information. In: *Proceedings of Workshop on Natural Language-Independent Engineering, 3rd Indian International Conference on Artificial Intelligence (IICAI-2007)*, Mumbay, pp.1814–1823.

Buckwalter T. (2002) *Buckwalter Arabic Morphological Analyzer Version 1.0* Linguistic Data Consortium, University of Pennsylvania.

Buckwalter, T. (2004). *Buckwalter Arabic morphological analyzer (BAMA) version 2.0.* linguistic data consortium (LDC) catalogue number LDC2004L02. ISBN1-58563-324-0.

Diab, M. (2009) Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging and Base Phrase Chunking. *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, 2009.

Diab, M, Hacioglu, K., & Jurafsky, D. (2007) Arabic Computational Morphology: Knowledge-based and Empirical Methods, chapter Automated Methods for Processing Arabic Text: *From Tokenization to Base Phrase Chunking.Kluwer*/springer edition

Farghaly, A., & Shaalan, K. (2009) Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP),* pp.1–22.

Goweder, A., Poesio, M., De Roeck, A. N., & Reynolds, J. (2004). Identifying Broken Plurals in Unvowelised Arabic Tex. In *EMNLP* (pp. 246-253).

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. *In COLING* (Vol. 96, pp. 466-471).

Habash N. (2010) *Introduction to Arabic Natural Language Processing*. Synthesis Lecture on Human Language Technologies. A Publication in the Morgan & Claypool Publishers series, UAS.

Habash, N., Rambow, O., & Roth, R. (2009) MADA+TOKAN: A toolkit for Arabic tokenization diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt

Khoja, S. (1999) *Stemming Arabic Text*. Computing Department, Lancaster University, Lancaster, U.K

Larkey, S., & Connell, E. (2001) Arabic Information Retrieval at *UMass In TREC-10, The Tenth Text Retrieval Conference, TREC 2001*. Gaithersburg: NIST, 562-570

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. *In Computational Linguistics and Intelligent Text Processing* (pp. 171-189). Springer Berlin Heidelberg.

Mohamed, E., & Kübler, S. (2010) Arabic Part of Speech Tagging. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LRE 2010C)*, 19-21 May, Valletta, Malta.

Peng, F., Feng, F., & McCallum, A. (2004). Chinese segmentation and new word detection using conditional random fields. *In Proceedings of the 20th international conference on Computational Linguistics* (p. 562). Association for Computational Linguistics.

Sawalha, M. & Atwell, E. (2009) Linguistically Informed and Corpus Informed Morphological Analysis of Arabic. In: *Proceedings of the 5th International Corpus Linguistics Conference CL2009*, 20-23 July 2009, Liverpool, UK.

Sawalha, M., & Atwell, E. (2008). Comparative evaluation of arabic language morphological analysers and stemmers. *In Proceedings of COLING 2008 22nd International Conference on Comptational Linguistics (Poster Volume))* (pp. 107-110). Coling 2008 Organizing Committee.

Voutilainen, A. (2003) Part-of-speech tagging. In R. Mitkov, editor, *The Oxford handbook of computational linguistics*. University Press, Oxford, pp. 219–232.