APT: Arabic Part-of-speech Tagger

Shereen KHOJA
Computing Department, Lancaster University
Lancaster
LA1 4YR, UK,
s.khoja@lancaster.ac.uk

Abstract

Arabic is the language of millions of people all over the world, yet a publicly available grammatically tagged corpus of Arabic still does not exist. In this paper, I describe some of the initial findings in the development of an Arabic part-of-speech tagger. For this tagger I have compiled a tagset containing 131 tags that is derived from traditional Arabic grammatical theory. I have used this tagger to manually tag a corpus and I have extracted a lexicon from this corpus. Because Arabic is a morphologically complex language, some pre-processing is necessary before automatic tagging can take place. The results I have obtained so far highlight some of the characteristics of the Arabic language that need to be addressed to improve tagging.

Introduction

Arabic is the official language of twenty Middle East and African countries, and is the religious language of all Muslims, regardless of their origin. It is therefore surprising that very little work has been done on Arabic corpus linguistics.

Arabic differs from Indo-European languages syntactically, morphologically and semantically. It is a Semitic language whose main characteristic feature is that most words are built up from roots by following certain fixed patterns[1] and adding infixes, prefixes and suffixes. It is an old language, and what is now known as Classical Arabic was standardised around fourteen centuries ago.

The modern form of Arabic is called Modern Standard Arabic (MSA) and it is the form used by all Arabic-speaking countries in publications, the media and academic institutions. MSA is spoken by people from different Arab countries where the local dialect may not be mutually intelligible. MSA is a simplified form of Classical Arabic, and follows its grammar. The main differences between Classical and MSA are that MSA has a larger (more modern) vocabulary, and does not use some of the more complicated forms of grammar found in Classical Arabic.

Statistical and hybrid part-of-speech taggers have been used very successfully for English (Garside, 1997) (Church, 1988), and with varying degrees of success for other European languages (Sanchez, 1995), (Tzoukermann, 1995). I am interested in seeing how these techniques can be applied to a language with a radically different morphology and syntax, and what results can be obtained.

In the following sections I will describe an Arabic part-of-speech tagger that I have called APT that uses statistical and rule-based techniques. I will also describe the tagset that I have derived from traditional Arabic grammatical theory, and that is still used today in Modern Standard Arabic. Finally, I will show some of the results that I have obtained so far.

1 Tagging Techniques

Many techniques have been used to tag English and other European language corpora. The first technique to be developed was the rule-based technique used by Greene and Rubin in 1970 to tag the Brown corpus (Greene, 1971). Their tagger (called TAGGIT) used context-frame rules to select the appropriate tag for each word. It achieved an accuracy of 77%. More recently, interest in rule-based taggers has re-emerged with Eric Brill's tagger that achieved an accuracy of 96% (Brill, 1992). The accuracy of this tagger was later improved to 97.5% (Brill, 1994).

In the 1980s more research effort went into taggers that used hidden Markov models to select

[1] Also called measures or *binyan*.

the appropriate tag. Such taggers include CLAWS, which was developed at Lancaster University and achieved an accuracy of 97% (Garside, 1987) (Garside, 1997), Church's PARTS tagger (Church, 1988) and the Xerox tagger, which was developed by Doug Cutting and achieved an accuracy of 96% (Cutting, 1992).

A combination of both statistical and rule-based methods has also been used to develop hybrid taggers. These seem to produce a higher rate of accuracy. An accuracy of 98% has been reported by Tapanainen and Voultilainen (1994) when using both techniques separately, then aligning the output. In fact, both Brill's tagger and CLAWS are in essence a combination of both techniques. In Brill's tagger, rules are selected that have the maximum net improvement rate that is based on statistics, and CLAWS contains a rule-based component that handles idioms.

More recently, taggers that use artificial intelligence techniques have been developed. One such tagger (Daelemans, 1996) uses machine learning, and is a form of supervised learning based on similarity-based reasoning. The accuracy rates of this tagger for English reached 97%. Also, neural networks have been used in developing Part-of-Speech taggers. The tagger developed at the University of Memphis (Olde, 1999) is an example of such a tagger. It achieved an accuracy of 91.6%. Another neural network tagger developed for Portuguese achieved an accuracy of 96% (Marques, 1996).

2 Arabic Tagset

Since the grammar of Arabic has been standardised for centuries, I decided to derive my initial tagset from this grammatical tradition rather than from an Indo-European based tagset. The reason for this is that Arabic is a very different language from Indo-European languages, and should have its own tagset. Also, Arabic linguists will be basing their studies on a traditional Arabic grammar rather than an Indo-European grammar.

Arabic grammarians traditionally analyse all Arabic words into three main parts-of-speech. These parts-of-speech are further sub-categorised into more detailed parts-of-speech which collectively cover the whole of the Arabic language (Haywood, 1962). The three main parts-of-speech are:

1. **Noun**: A noun in Arabic is a name or a word that describes a person, thing, or idea. Traditionally the Noun class in Arabic is subdivided into Derivatives (that is, nouns derived from verbs, nouns derived from other nouns, and nouns derived from particles) and Primitives (nouns not so derived). These nouns could be further sub-categorised by number, gender and case. This class also includes what, in traditional European grammatical theory, would be classified as Participles, Pronouns, Relatives, Demonstratives and Interrogatives.
2. **Verb**: The verb classification in Arabic is similar to that in English, although the tenses and aspects are different. The Verb tag can be sub-categorised into Perfect, Imperfect, and Imperative. Further sub-categorisation of the Verb class are possible using number, person and gender.
3. **Particle**: The Particle class includes: Prepositions, Adverbs, Conjunctions, Interrogative Particles, Exceptions[2], and Interjections.

Initial experiments used a tagset that had five main tags (Noun, Verb, Particle, Residual, Punctuation), which needed to be extended to 35 tags taking into account clitics (see section 3). The tagset has now been revised and contains 131 tags. These are all subcategories of the five main categories.

The verbs have been sub-categorised by 'type' (perfect, imperfect, imperative), person, number and gender, and the tag name reflects this sub-categorisation. For example, the word *ksrtm*[3] "you [plural, masculine] broke", which is a perfect verb in the second person masculine plural form, has the tag VPPl2M. An indicative imperfect second person feminine singular verb such as *tktbyn* "you [singular, feminine] are writing" would be tagged VISg2FI.

Similarly, personal pronouns are tagged for number, person and gender. The tag for *hma* that is a personal pronoun for the third person dual neutral (masculine or feminine) is tagged as NPrPDu3. As well as personal pronouns, there are

[2] These include the Arabic words that are equivalent to the word except and the prefixes non-, un-, and im-.

[3] Words in italic are tranliterations of the Arabic word.

relative and demonstrative pronouns, which are also classed by number and gender.

Nouns are classed by number (compare *ktab*, *ktaban*, *ktb*, meaning "one book", "two books", and "books") and gender (compare *ktab* [masculine] meaning "book" and *mdrst* [feminine] meaning "school")). Foreign and proper nouns receive separate tags.

The category of particle includes prepositions, adverbs and conjunctions, all of which appear in Arabic either as individual words, or as clitics attached to the following word (see section 3). Other particles are interjections, exceptions and negative particles.

Dates, numbers, punctuations and abbreviations are also tagged separately. Dates are classified as either an Arabic date such as *mhrm* which is the first month in the Islamic calendar, or a date from the Gregorian calendar such as *a<u>t</u>ar* which is the Arabic word for the month "March"[4].

3 The Training Corpus and Lexicon

A corpus of 50,000 words in Modern Standard Arabic (an extract from the Saudi Al-Jazirah newspaper, date 03/03/1999) was tagged using the smaller tagset described above, and is currently being retagged with the more detailed tagset.

For morphologically complex words a combination of tags was used. For example, the word *walktab* "and the book" is given the tag PC+NCSgMND, where PC indicates a particle that is a conjunction, and NCSgMND indicates a singular, masculine, nominative, definite noun.

This corpus was used to derive the various lexicons described below. It was also used to train the tagger; that is, to extract the statistical data that is needed for the automatic tagging of untagged or raw corpora (see section 4.4).

The first version of the lexicon lists every word that appears in the training corpus together with all the tags it has received. Because of the frequent occurrence of clitics in Arabic, many words appear in the lexicon several times with different clitics. For example, the word *mst<u>s</u>fat* "hospital" appears twice in the lexicon, once with the definite article, and once without.

An example of a clitic in English would be "she'll", where two words have been merged together, "she" and "will". In Arabic the definite article, equivalent to "the" in English, appears as a two-letter proclitic at the beginning of the noun.

This is similar to the definite article in French that appears as a proclitic when attached to a word that starts with a vowel. An example is *l'idée*, where *l'* is the definite article, and *idée* is the French word for "idea". Another example of a clitic is the Italian word *muoviamoci* "let us go". The clitic here is the first person plural pronoun *ci*.

Similarly the conjunction "and" appears in Arabic as a one-letter proclitic attached to the first word of the coordinated sequence; this word could be a noun (including a noun with prefixed definite article), a verb, particle or even a number. Other Arabic clitics include further inseparable conjunctions, pronouns, and some prepositions.

For the second version of the lexicon, the clitics are removed from the words before they are placed in the lexicon. This allows a rule-based component of the tagger to match the un-cliticised word in the lexicon to the word in the running text, whatever clitics it is encountered with. In the original lexicon derived from the training corpus, for instance, the ambiguous Arabic word *mars* appears without any clitics meaning the foreign month name "March", and also with the conjunction "and" to mean the verb "and he practiced".

The initial version of the lexicon containing all words from the corpus without removing any clitics contained 13,912 words. After removing clitics, the lexicon was reduced to 9,986 words. Table 1 shows an extract from the lexicon. The first column contains a transcription of the Arabic word, the second column contains its meaning, and the other columns contain all that words tags.

This small extract from the lexicon highlights the fact that the lexicon does not cover all the possible tags of some words. For example, the Arabic word *<u>h</u>rs* appears in the corpus only meaning the noun "guards", but more generally it could also be the verb "he guarded". Another example is the word *<u>h</u>sn* that appears three times in the training corpus as a proper noun, but it does not appear in the training corpus with its other meaning "good". It will be necessary in the future to go through the lexicon and manually add all the possible tags to each word, or use a larger training

[4] Some Arabic texts use a transcription of the English months. For example, March transcribed in Arabic as *mars*.

corpus, thought that might still not find all possible tags.

thml	carrying, carry	VISG2M	VISG3F	
twjyhat	instructions	NPLF		
kadm	servant	NSGM		
?qd	convene, contract	VPSG3M	NSGM	
hrmyn	mosques [dual]	NDUM		
sryfyn	distinguished [dual]	NDUM		
mustsfa	hospital	NSGF		
an	truly, not, letter N	P	PS	NF
mtnql	mobile	NSGM		
adwyt	medicines	NPLF		
ila	to	PPS		
ahd	one, Sunday	NSGM	ND	
nql	move(V), move(N)	VPSG3M	NSGM	
?ly	Ali(name), on	NP	PPS	
t?ml	doing	VISG2M	VISG3F	
hrs	guard	NPLM		
hsn	Hassan(name)	NP		

Table 1: Extract from the lexicon

4 The Arabic POS Tagger

APT was developed using a combination of both statistical and rule-based techniques since hybrid taggers seem to produce the highest accuracy rates (see section 1).

4.1 Initial Tagging

The first step of the tagger is the initial tagging. This is basically the look-up component of the tagger. Here every word is looked up in the lexicon, and if it is found, then it is given all the possible tags of that word as specified in the lexicon.

In early versions of the tagger, clitics would be stripped off words in the running text before look-up in the lexicon, but this is now incorporated in the more general morphological analysis described in the section 4.2.

Since the lexicon is so small, the initial look-up is unlikely to find many of the words. Most of the words that it would find might be particles and if the text is from the same genre, then maybe some of the proper nouns and verbs.

It is obvious then that because of Arabic's complex morphology, some pre-processing or morphological analysis is required. This is achieved by using the stemmer.

4.2 Stemming

Stemming is the process of removing all of a word's affixes to produce the stem or root. In Arabic this means the removal of prefixes, suffixes and infixes. The stemming component is the rule-based part of the tagger, since rules are used to determine what the affixes are.

After the initial tagging, if a word is not found in the lexicon, it is stemmed. The affixes are used to help determine the tag of the word. Sometimes one affix can determine the tag of a word: for example, if the prefix is the definite article, then the word is a noun. On the other hand, a combination of affixes is most commonly used to determine the tag of the word. For example, if the prefix is a *t* "which indicates the imperfect verb" and the suffix is *wn* "which indicates the masculine plural", then the word is likely to be a second person plural masculine imperfect verb, such as *tdrswn* which means "you [plural masculine] are studying".

Since the stemming algorithm also uses the Arabic word patterns, these can be used to determine the tag of the word. Most words in Arabic are formed using fixed patterns, and these patterns have predictable properties and meanings. For example, words that follow a certain pattern are plural nouns. This type of plural is called the broken plural, and it is different from the sound or perfect plural because it is not formed by the simple addition of suffixes.

Some of the problems that the stemmer faces are that some letters that appear to be affixes are in fact part of the word. Another problem is that some letters (for instance the long vowels) may change to other letters when an affix is added, and so the letters should be changed back when that affix is removed.

Tests of the stemmer over Arabic words show that it achieves an accuracy of 97%, using a dictionary of 4,748 triliteral and quadriliteral roots.

4.3 Results after Initial Tagging

Four corpora were compiled and used for testing. The corpora were chosen because they are from different countries, and although they all use Modern Standard Arabic, some local colloquial differences do appear. The last corpus was chosen

because it is from a different genre. The corpora are:

- The remaining 59,040 words of the Saudi `` Al-Jazirah" newspaper, dated 03/03/1999.
- 3,104 words of the Egyptian `` Al-Ahram" newspaper, date 25/01/2000.
- 5,811 words of the Qatari `` Al-Bayan" newspaper, date 25/01/2000.
- 17,204 words of Al-Mishkat, an Egyptian published paper in social science, April 1999.

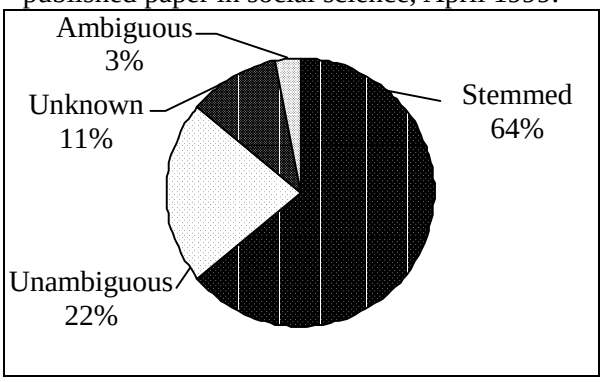

Figure 1: Statistics from Al-Jazirah

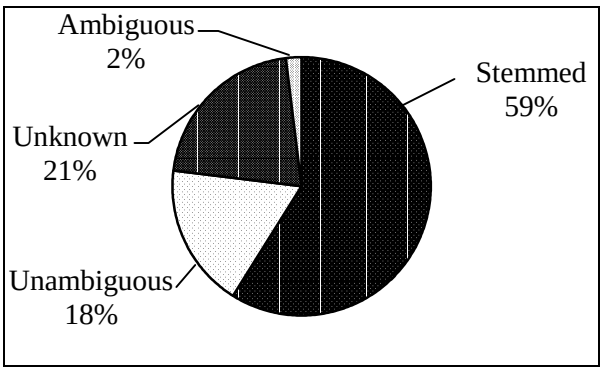

Figure 2: Statistics from Al-Ahram

The results of the initial tagging for these corpora are shown in Figures 1-4.

These initial results show that the number of words that were found in the lexicon range from 20% in Al-Ahram to 26% in Al-Bayan. On the other hand only a small number of these (1-3%) were ambiguous. The true ambiguity rate is likely to be higher when (a) all possible tags have been added to the lexicon, and (b) the results of the stemmer have been analysed, since most words in the stemmer are expected to be ambiguous.

The results also show that a very high number of words in the test corpora (10-21%) are not found in the lexicon, and could not be handled correctly by the stemmer (unknown words). The reason that the stemmer could not handle these words is usually that they are foreign words or proper nouns.

Another interesting point that we note here is that the frequency of unknown words in both Al-Ahram and Al-Mishkat are significantly higher than in the other two corpora. Both these corpora are Egyptian, and although they were written in Modern Standard Arabic, and so should not be significantly different, the vocabulary used might be slightly different. Also, these corpora are not free from dialectal words that tend to creep into MSA.

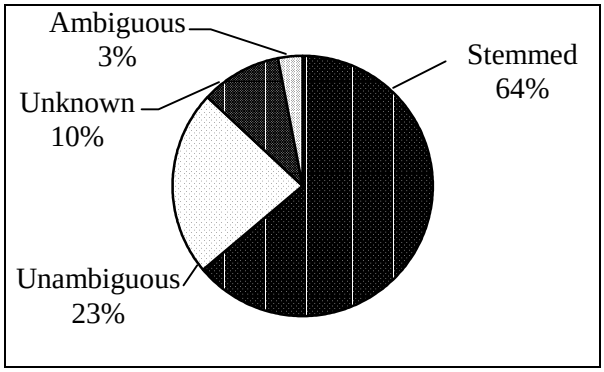

Figure 3: Statistics from Al-Bayan

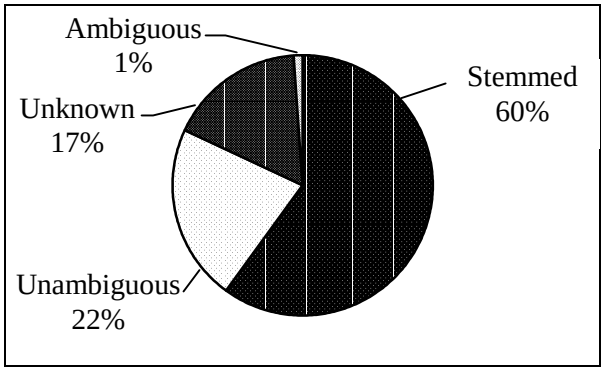

Figure 4: Statistics from Al-Mishkat

4.4 Tag Disambiguation

A statistical tagger that uses the Viterbi algorithm (Jelinek, 1976) has been developed and used to disambiguate words that have more than one tag (ambiguous words and unknown words).

The probabilities used by the tagger are calculated from the tagged training corpus described in section 3. The two probabilities used are the *lexical probability*, which is the probability of a word having a certain tag, and the *contextual probability*, which is the probability of one tag following another tag.

Probabilities have been derived for the smaller tagset described in section 2. The contextual

probabilities that were calculated can be seen in Table 2. Here, the probability of a Noun being followed by another Noun is 0.711, and the probability of a Verb being followed by a Noun is 0.926.

	N	V	P	No.	Pu.
N	0.711	0.065	0.143	0.010	0.071
V	0.926	0.037	0.0	0.008	0.029
P	0.689	0.199	0.085	0.016	0.011
No.	0.509	0.06	0.098	0.009	0.324
Pu.	0.492	0.159	0.152	0.046	0.151

Table 2: Contextual Probabilities of Tagged Corpus

The statistical tagger achieved an accuracy of around 90% when disambiguating *ambiguous words* with this tagset. On the other hand, the tagger tagged all *unknown words* as nouns. The reason for this is that over 30,000 words (67%) in the training corpus were nouns, so the probability of a word being a noun is much higher than the probability of it being anything else.

Conclusion

I am currently recalculating the lexical and contextual probabilities that will be used by the tagger for the larger tagset. The results of the statistical tagger will be presented in the near future.

Another improvement that is necessary here is to go through the lexicon manually and add all the possible tags that a word can take.

A component that still needs to be added is a pre-processing component to handle common errors found in Arabic. These include the errors in the placement of the *hamza* on the *alif* (glottal stops), and also placing the dots under the letter yah. An example of such an error can be seen in Table 1, where the name *ali* should have the dots, while the preposition meaning "on" should not have the dots.

References

Roger Garside, Geoffrey Leech, and Anthony McEnery (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora.* Addison Wesley Longman Inc., New York.

Kenneth Church (1988) *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text.* In "Proceedings of the Second Conference on Applied Natural Language Processing" (ACL) Austin, Texes, pp. 136-143.

Fernando Sanchez Leon and Amalio F. Nieto Serrano. (1995) *Public Domain POS Tagger for Spanish.* CRATER Final Deliverable Report No. 22.

Evelyne Tzoukermann, Dragomir R. Radev, and Willian A. Gale (1995) *Combining Linguistic Knowledge and Statistical Learning in French Part-of-Speech Tagging.* In EACL SIGDAT Workshop. Association for Computational Linguistics - European Chapter, Dublin, Ireland.

B.B. Greene and G.M. Rubin (1971) *Automatic Grammatical Tagging of English.* Department of Linguistics, Brown University, Providence, R.I.

Eric Brill (1992) *A Simple Rule-Based Part of Speech Tagger.* In "Proceedings of the Third Conference on Applied Natural Language Processing", Trento, Italy, pp. 152-155.

Eric Brill (1994) *Some Advances in Transformation Based Part of Speech Tagging*. In "Proceedings of the Twelfth International Conference on Artificial Intelligence" (AAAI-94), Seattle, WA.

Roger Garside, Geoffrey Leech, and Geoffrey Sampson (1987) *The Computational Analysis of English: a corpus-based approach*. Longman Group UK Limited.

Doug Cutting, Julian Kupiec, Jan Pederson, and Penelope Sibun (1992) *A Practical Part-of-Speech Tagger.* In "Proceedings of the Third Conference on Applied Natural Language Processing", Trento, Italy.

Pasi Tapanainen and Atro Voutilainen (1994) *Tagging accurately - Don't guess if you know.* In "Proceedings of the Fourth ACL Conference on Applied Natural Language Processing", Stuttgart.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis (1996) *MBT: A Memory-Based Part of Speech Tagger-Generator.* In "Proceedings of the Fourth Workshop on Very Large Corpora", Copenhagen, Denmark, pp. 14-27.

Brent A. Olde, James Hoener, Patrick Chipman, Arthur C. Graesser, and the Tutoring Research Group (1999) *A Connectionist Model for Part of Speech Tagging.* In "Proceedings of the 12th International Florida Artificial Intelligence Research Society Conference", Menlo Park, CA, pp. 172-176.

Nuno Marques and Jos Gabriel Lopes (1996) *Using Neural Nets for Portuguese Part-of-Speech Tagging.* In "Proceedings of the Fifth International Conference on The Cognitive Science of Natural Language Processing", Dublin City University.

JA Haywood and H M Nahmad (1962) *A new Arabic grammar.* Lund Humphries Publishers Ltd.

F. Jelinek (1976) *Continuous Speech Recognition by Statistical Methods.* In "Proceedings of the IEEEE", pp. 532-556.