Developing a New System for Arabic Morphological Analysis and Generation

Mourad Gridach
Mathematics and Computer Science Department Faculty of Science Dhar El Mehraz Fez
Mourad_i4@yahoo.fr

Noureddine Chenfour
Mathematics and Computer Science Department Faculty of Science Dhar El Mehraz Fez
chenfour@yahoo.fr

Abstract

Arabic morphology poses special challenges to computational natural language processing systems. Its rich morphology and the highly complex word formation process of roots and patterns make computational approaches to Arabic very challenging. In this paper we present an approach for morphological analysis and generation of Modern Standard Arabic (MSA). Our approach is based on Arabic morphological automaton technology. We take the special representation of Arabic morphology (root and scheme) to construct a set of morphological automaton which will be used directly in developing a system for Arabic morphological analysis and generation. Our approach for Arabic morphological analysis and generation can be used in different Arabic NLP applications such as Machine Translation (MT) and Information Retrieval (IR).

1 Introduction

Due to the rising importance of globalization and multilingualism, there is a need to build natural language processing (NLP) systems for an increasingly wider range of languages, including those languages that have traditionally not been the focus of NLP research. The development of NLP technologies for a new language is a challenging task since one needs to deal not only with language specific phenomena but also with a potential lack of available resources (e.g. lexicons, text, annotations).

Arabic is a language of rich morphology compared to other language especially European languages. It based on both derivational and inflectional morphology. The richness of Arabic morphology makes the analysis process difficult to deal. On the one hand, morphological analysis process is used in the most of the NLP applications such as information retrieval, spell-checking and machine translation. On the other hand, morphological analysis is the first step before syntactic analysis. Furthermore, it is an essential step in semantic analysis.

There has been much work on Arabic morphology. For an overview see (Al-Sughaiyer and Al-Kharashi, 2004). Generally speaking, morphological analysis of any word given consists of determining the values of a large number of features such as basic part-of-speech (i.e., noun, verb, etc.), gender, person, number, voice, information about the clitics, etc. (Habash, 2005). The most of the morphological analysis systems don't display the whole features of the word analyzed and some of them are destined for a special applications. We note that the morphological analysis systems available now have different aims, some of them have a commercial purpose and the other systems are available for research and evaluation (Attia, 2006).

In this paper we present an approach for Arabic morphological analysis and generation based on morphological automata and used a morphological database constructed using XMODEL (XML-base Morphological Definition Language). To develop an Arabic morphological automaton, we exploited particularities of Arabic morphology. The Arabic verbs and nouns are characterized by a special representation "root + scheme". Verbs and nouns are derived from roots by applying schemes to these roots to generate Arabic stems and then adding prefixes and suffixes to the stems to form a correct word in Arabic language. Table **1** show four schemes applied to the root "cml" (the work notion) (عمل) to generate four derived stems.

Scheme	facal	FAcil	fuccAl	Mafcal
Stem generated	عَمَل	عَامِل	عُمَّال	مَعْمَل
Transliteration	camal	CAmil	cummAl	macmal

Table 1 : Schemes generating stems from the root "cml" (عمل)

Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP 2011, , pages 52–57, Chiang Mai, Thailand, November 8, 2011.

2 Previous work

There has much been work on Arabic morphological analysis and generation. In this paragraph, we will present some of the most work referenced in the literature and well documented.

2.1 ElixirFM: an Arabic Morphological Analyzer by Otakar Smrz

ElixirFM is an online Arabic Morphological Analyzer for Modern Written Arabic developed by Otakar Smrz available for evaluation and well documented. This morphological analyzer is written in Haskell, while the interfaces in Perl. ElixirFM is inspired by the methodology of Functional Morphology (Forsberg & Ranta, 2004) and initially relied on the re-processed Buckwalter lexicon (Buckwalter, 2002). It contains two main components: a multi- purpose programming library and a linguistically morphological lexicon (Smrz, 2007). The advantage of this analyzer is that it gives to the user four different modes of operation (Resolve, Inflect, Derive and Lookup) for analyzing an Arabic word or text. But the system is limited coverage because it analyzes only words in the Modern Written Arabic.

2.2 MAGEAD: A Morphological Analyzer and Generator for Arabic Dialects

MAGEAD is one of the existing morphological analyzers for the Arabic language available for research. It's a functional morphology systems compared to Buckwalter morphological analyzer which models form-based morphology (M. Altantawy et al., 2010). To develop MAGEAD, they use a morphemic representation for all morphemes and explicitly define morphophonemic and orthographic rules to derive the allomorphs. The lexicon is developed by extending Elixir-FM's lexicon. The advantage of this analyzer is that it processes words from the morphology of the dialects which they considered as a novel work in this domain, but unfortunately this analyzer needs a complete lexicon for the dialects to make the evaluation more interesting and convincing, and to verify these claims.

2.3 Buckwalter Arabic Morphological Analyzer

This analyzer is considered as one of the most referenced in the literature, well documented and available for evaluation. It is also used by Linguistic Data Consortium (LDC) for POS tagging of Arabic texts, Penn Arabic Treebank, and the Prague Arabic Dependency Treebank (Atwell et al., 2004). It takes the stem as the base form and root information is provided. This analyzer contains over 77800 stem entries which represent 45000 lexical items. However, the number of lexical items and stems makes the lexicon voluminous and as result the process of analyzing an Arabic text becomes long.

2.4 Xerox Arabic Morphological Analysis and Generation

Xerox Arabic morphological Analyzer is well known in the literature and available for evaluation and well documented. This analyzer is constructed using Finite State Technology (FST) (Beesley, 1996; Beesley, 2000). It adopts the root and pattern approach. Besides this, it includes 4930 roots and 400 patterns, effectively generating 90000 stems. The advantages of this analyzer are, on the one hand, the ability of a large coverage. On the other hand, it is based on rules and also provides an English glossary for each word. But the system fails because of some problems such as the overgeneration in word derivation, production of words that do not exist in the traditional Arabic dictionaries (Darwish, 2002) and we can consider the volume of the lexicon as another disadvantage of this analyzer which could affect the analysis process.

3 Our approach

3.1 Lexicon

The lexicon of a language is the set of its valid lexical forms. As in any morphological analysis system, developing a high-quality lexicon is often the first step towards building a robust morphological analyzer, which is in turn the front-end to many NLP systems. There are two aspects that contribute to this enhancement level. The first aspect concerns the number of lexicon entries contained in the lexicon. Second aspect concerns the richness in linguistics information contained by the lexicon entries. BAMA lexicon is the best know in the literature and well documented. It used by large Arabic morphological analyzers (Elixir-FM and MAGEAD).For an overview of the existing Arabic lexicon see (Al-Sughaiyer and Al-Kharashi, 2004).

Nowadays, a new method was been implemented to represent, design and implement the lexicons. It is based on the Lexical Markup

Framework (LMF). LMF is the ISO-24613 standard for natural language processing (NLP) and lexicons. The US delegation is the first which started the work on LMF in 2003. In early 2004, the ISO/TC37 committee decided to form a common ISO project with Nicoletta Calzolari (Italy) as convenor and Gil Francopoulo (France) and Monte George (US) as editors. The aims of LMF are to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources. This method for representing lexical resource covers all the natural languages. We note that for Arabic language, lexicons based on LMF are still in progress towards a standard for representing the Arabic linguistic resource.

Our approach for representing the lexicon is based on XMODEL (XML-based Morphological Definition Language). In this approach, the Arabic lexicon contains morphological classes, morphological properties and morphological rules. Morphological classes allow gathering a set of morphological components having the same nature, the same morphological characteristics and the same semantic actions. For the morphological properties, they allow characterizing the different morphological components represented by the morphological classes; they contain morphological descriptors (the features) that would be assigned to different morphological components (the property "Gender" distinguishes between masculine and feminine components). Finally, morphological rules allow combining the morphological components to generate correct language words. They are considered as a generator of language words. We note that until now, our morphological database contains 5970 entries. The use of XMODEL allows representing the morphological database independent of processing which will be applied and allows a considerable reduction of morphological entries.

3.2 System description

In this part we describe the Arabic morphological analyzer. So as to develop this analyzer, first of all, we developed an Arabic morphological database using XMODEL language integrating all the entries suitable for Arabic language. Then, we generated a set of Arabic morphological automata representing a specific morphological category. Finally, a framework is developed to handle the lexicon and the morphological automata.

The presented work involves five steps. In this paragraph, we provide a brief description of the principles of this work. As input, the proposed technique accepts an Arabic text. The first step is to apply a tokenization process to the text given. Then, a set of AMAUT (Arabic Morphological AUTomata) are loaded, in a second step. The part-of-speech is determined in the third step. After that, the method determines all possible affixes. Then the next step consists of extracting the morpho-syntactic features according to the valid affixes.

The tokenization process consists of extracting all the words from the text given. A set of Arabic morphological automata are loaded from a package that contains all the implemented Arabic morphological automata. Then, the approach determines which AMAUT is suitable for that word. The result may be one or more AMAUT loaded. We note that the size of the final AMAUT generated is about 120 MB. Then, the method determines the part-of-speech. If the word analyzed is a noun or a verb, the method determines if it contains a scheme. Then, if it is a verb, the method determines the type of the verb (strong, weak, or incomplete), its tense ("mADI" /ماضي/, "muDAric" /مضارع/ or "eamr" /أمر/), its voice (active or passive), etc. If it is a noun, we determine if it is a derived noun or particular noun. If it is a particle, the method determines if it is a preposition particle /حروف الجر/, conjunction particle /حروف العطف/ , etc. After that, the method applied a process of extracting the possible affixes attached to the word analyzed. The next step consists of extracting the morpho-syntactic features according to the valid affixes and the scheme. Additional information is extracted called in our approach morphological descriptors. They describe the word analyzed and they are very useful especially in Natural Language Processing applications. Finally, the morphological analyzer displays the results in a table where each row contains the word analyzed and all the data characterizing this word (see Figure 1).

Generally speaking, morpho-syntactic features displayed by the morphological analyzer are very rich regarding the information given. It concerns the morphological level; the syntactic and semantic level which makes the richness of our system compared to the others system. The utility of this richness comes especially when the system will be used in NLP applications. Here

are the most important features given by the system.

- The word gender: masculine or feminine.
- The word person: first, second or third person.
- The word number: singular, dual or plural.
- The word case: “marfUc” (مرفوع), “manSUb” (منصوب), “majrUr” (مجرور), “majzUm” (مجزوم).
- The type of the word: verb, noun or particle.
- If the word is a verb, we give its tense: present (“ealmuDAric”: المضارع), past (“ealmADI”: الماضي) or imperative (“ealeamr”: الأمر). We also give its voice: active or passive.
- The scheme of the word is given if available.

Figure **1** shows the morphological analysis results of some words analyzed using the presented morphological analyzer. The displays the Part-of-speech (verb, noun or particle), the original scheme is displayed in column B because Arabic has this particularity which is summarized in that some words might be conjugated forms of other words like “afcalu”, “afcilu “, “afculu”, these three words are all conjugated forms of “facala”. The gender (masculine or feminine) is displayed in column D, the person (first, second or third person) is displayed in column E, the number (singular, dual or plural) is displayed in column F. For the column G, it concerns some properties that characterize the word analyzed and they are very useful to the user. Some morphological descriptors are displayed in column H. Finally, the column I and J show the affixes attached to the word.

Lexical Table of File : arab files\dfa01.txt

A	B	C	D	E	F	G	H	I	J
Morphological	Original Scheme	Scheme	Gender	Person	Number	Properties	Morphological Descriptors	Prefixes	Suffixes
yatadaHrajAni	[tcfaclala],	[]	GMa	Fr3	NDl	Strong Verb,MOD,ACT,	Raf,	[y]	[Ani]
eaSTaffa	[efcalla],	[]	GFe,GMa	Fr1	NSg	Strong Verb,ACT,MOD,	Def,NaS	[e]	[a]
eaSTaffa	[efcalla],	[]	GFe,GMa	Fr1	NSg	Strong Verb,ACT,MOD,	Def,NaS	[e]	[a]
eaSTaffu	[efcalla],	[]	GFe,GMa	Fr1	NSg	Strong Verb,ACT,MOD,	Def,Raf,	[e]	[u]
yaqudduu	[cadda],	[]	GMa	Fr3	NSg	Incomplete Verb,MOD,	Def,Raf,	[y]	[uu]
yadidu	[wacala, wacila, wacula],	[]	GMa	Fr3	NSg	Weak Verb,ACT,MOD,	Def,Raf,	[y]	[u]
yuda	[fcC\, faciya],	[]	GMa	Fr3	NSg	Weak Verb,PAS,MOD,	NaS,Jaz,	[y]	[a]
yar~a	[wacala, wacila, wacula],	[]	GMa	Fr3	NSg	Weak Verb,ACT,MOD,	Def,NaS	[y]	[a]
yar~i	[eafcA],	[]	GMa	Fr3	NSg	Weak Verb,ACT,MOD,	NaS,Jaz,	[y]	[i]
yur~a	[fcC\, faciya],	[]	GMa	Fr3	NSg	Weak Verb,PAS,MOD,	NaS,Jaz,	[y]	[a]
yar~u	[wacala, wacila, wacula],	[]	GMa	Fr3	NSg	Weak Verb,ACT,MOD,	Def,Raf,	[y]	[u]

Figure 1: A morphological analysis of some Arabic words using the presented system

It should be noted that the presented system could be used in both analysis and generation unlike some Arabic morphological analyzers which cannot be converted to generators in a straightforward manner (Cavalli-Sforza, 2000; Buckwalter, 2004; Habash, 2004 ;).

4 Evaluation

To evaluate our system, we select two of the best known morphological analyzers in the literature: ElixirFM by Otakar Smrž (Otakar Smrž and Viktor Bielický, 2010) and Xerox Arabic Morphological Analyzer. We note that the corpus used for the evaluation is taken from a standard input text provided by ALECSO (Arab League, Educational, Cultural and Scientific Organization) which organized a competition in April 2009 of the Arabic Morphological Analyzers in Damascus.

The evaluation process shows that our morphological analyzer is strong concerning the features given by each analyzer which makes our system useful for the most of NLP applications unlike the others; they are destined for specific applications. In addition, the presented morphological analyzer gives more additional information about each word analyzed and more precision.

In the evaluation done we process words in a corpus selected from ALECSO input text containing different part-of-speech (verbs, nouns and particles), then, we calculate accuracy of each analyzer as: S = number of words with good solutions / number of words. Table 2 provides the evaluation results of the three analyzers. Note that Table 2 contains in each column of the analyzers the number of words (nouns, verbs and particles) with no solution.

POS	The number	Xerox Morphological Analyzer	ElixirFM	Our System
Nouns	576	60	56	40
Verbs	457	31	24	19
Particles	167	42	45	-
Total	1200	133	125	59
Accuracy (%)		88.91%	89.58%	95.08%

Table 2: The evaluation process results

The analyzer presented in this paper reaches an accuracy of 95.08% which will make it one of the best existing morphological analyzers for Arabic language and it will be very useful for the next future works to be done in NLP applications such as syntactic and semantic analysis, machine translation, information retrieval, etc.

5 Conclusion

In this paper, we have discussed some previous work in this area of research which is the most referenced in the literature. Then, we have outlined some challenges of computational Arabic morphology. After that, we presented an approach to develop a morphological analyzer and generator for Arabic language. To develop this system for Arabic morphological analysis, the need to develop a lexicon is an essential stage. So, we used a new language for representing, designing and implementing the linguistic resource. It is based on a reduced XML lexicon and it can be used not only in morphological level, but in the other levels such as syntactic and semantic level. Finally, our approach could be used in NLP applications such as machine translation and information retrieval.

Appendix (1): Letter mappings

ا	:	A	س	:	S	ك	:	k
ب	:	B	ش	:	^	ل	:	l
ت	:	T	ص	:	S	م	:	m
ث	:	~	ض	:	D	ن	:	n
ج	:	J	ط	:	T	هـ	:	h
ح	:	H	ظ	:	Z	و	:	w
خ	:	X	ع	:	c	ي	:	y
د	:	D	غ	:	g	ى	:	A
ذ	:	V	ف	:	f	ة	:	t
ر	:	R	ق	:	q	ء	:	e
ز	:	Z						

References

Al-Sughaiyer Imad A. and Al-Kharashi Ibrahim A. 2004. Arabic morphological analysis techniques: A comprehensive survey. Journal of the American Society for Information Science and Technology, 55(3):189–213

Altantawy Mohamed, Nizar Habash, Owen Rambow, and Ibrahim Saleh. 2010. Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. In Proceedings of the Language Resource and Evaluation Conference (LREC-2010), Malta.

Attia, M. 2006. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. The Challenge of Arabic for NLP/MT Conference, the British Computer Society, London.

Atwell E., Al-Sulaiti L., Al-Osaimi S., Abu Shawar B.. 2004. A Review of Arabic Corpus Analysis Tools, JEP-TALN 04, Arabic Language Processing, Fès, 19-22 April.

Beesley KR 1996. Arabic Finite-State Morphological Analysis and Generation, Proceedings of the 16th conference on Computational linguistics, Vol 1. Copenhagen, Denmark: Association for Computational Linguistics, pp 89-94.

Beesley KR. 2000. Finite-State Non-Concatenative Morphotactics SIGPHON-2000, Proceedings of the Fifth Workshop of the ACL Special Interest Group in Computational Phonology, p. 1-12, August 6, 2000, Luxembourg.

Buckwalter T. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, University of Pennsylvania, LDC Catalog No.: LDC2002L49.

Buckwalter, T. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. Linguistic Data Consortium, catalog number LDC2004L02 and ISBN 1-58563-324-0.

Cavalli-Sforza, V., Soudi. A, and Teruko M. 2000. Arabic Morphology Generation Using a Concatenative Strategy. In Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2000), Seattle, USA.

Darwish K. (2002). Building a Shallow Morphological Analyzer in One Day, Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02). Philadelphia, PA, USA.

Habash Nizar. 2004. Large scale lexeme based Arabic orphological generation. In Proceedings of Traitement Automatique du Langage Naturel (TALN-04). Fez, Morocco.

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

Forsberg M. and Ranta A. 2004. Functional Morphology ICFP'04, Proceedings of the Ninth ACM SIGPLAN International Conference of Functional Programming, September 19-21, Snowbird, Utah

Ibrahim, K. 2002. Al-Murshid fi Qawa'id Al-Nahw wa Al-Sarf [The Guide in Syntax and Morphology Rules]. Amman, Jordan, Al-Ahliyyah for Publishing and Distribution.

Otakar Smrz (2007). ElixirFM. Implementation of Functional Arabic Morphology. In ACL Proceedings of the Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, pages 1–8, Prague, Czech Republic.

Otakar Smrž and Viktor Bielický. 2010. ElixirFM. Functional Arabic Morphology, http://sourceforge.net/projects/elixir-fm/.