Arabic morpho-syntactic feature disambiguation in a translation context

Ines Turki Khemakhem, Salma Jamoussi, Abdelmajid Ben Hamadou

MIRACL Laboratory, ISIM Sfax, Pôle Technologique

ines_turki@yahoo.fr, salma.jammoussi@isimsf.rnu.tn, abdelmajid.benhamadou@isimsf.rnu.tn

Abstract

Morphological analysis and disambiguation are crucial stages in a variety of natural language processing applications such as machine translation, especially when languages with complex morphology are concerned such as Arabic. Arabic is a highly flexional language, in that, the same root can lead to various forms according to its context. In this paper, we present a system which disambiguates the output of a morphological analyzer for Arabic. The Arabic morphological analyzer used consists of a set of all possible morphological analyses for each word, with the unique correct syntactic feature. We want to choose the correct features using the features generated by the morphological analyzer for the French language in the other side. To obtain this data, we used the results of the alignment of word trained with GIZA++ (Och and Ney, 2003).

1 Introduction

Arabic is characterized by a rich morphology. Due to the fact that the Arabic script usually does not encode short vowels, the degree of morphological ambiguity is very high. In addition to being inflected for gender, number, words can be attached to various clitics for conjunction "و" (and), the definite article "ال" (the), prepositions (e.g. "ب" (by/with), "ل" (for), "ك" (as)), object pronouns (e.g. "هم" (their/them)).

The morphological analysis of a word consists of determining morphological information about each word, such as part-of-speech (i.e., noun, verb, particle, etc), voice, gender, number, information about the clitics, etc. Morphological analysis and disambiguation are crucial preprocessing steps for a variety of natural language processing applications, from search and information extraction to machine translation. For languages with complex morphology these are nontrivial processes.

Arabic words are often ambiguous in their morphological analysis. This is due to Arabic's rich system of affixation and clitics and the omission of disambiguating short vowels. The problem is that many words have different meanings depending on their diacritization. This leads to ambiguity when processing data for natural language processing applications such as machine translation. This propriety has an important implication for statistical modeling of the Arabic language.

In this paper, we present a novel morphology preprocessing technique for Arabic. We exploit the Arabic morphology-French alignment to choose a correct morpho-syntactic feature produced by a morphological analyzer.

This paper is organized as follows: section 2 gives a brief description of some related works to the introduction of morphological disambiguation. Section 3 presents the used morphological analyzer Morph2 for Arabic texts, able to recognize word composition and to provide more specific morphological information about it. We present in section 4 some problems in context morpho-syntactic feature choice; in the remainder of this section we discuss the complexity of Arabic morphology and the challenge of morphological disambiguation. Section 5 gives a short overview of the data and tools used for our Arabic word Morpho-syntactic feature disambiguation and shows the experimental details of our system. Finally, section 6 presents some conclusions.

Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation, pages 61–65,
COLING 2010, Beijing, August 2010.

2 Related work

Morphological analysis and disambiguation are crucial pre-processing steps for a variety of natural language processing applications.

Previous research has focused on disambiguating the output of a morphological analyzer. Hajic (2000) is the first to use a dictionary as a source of possible morphological analyses (and hence tags) for an inflected word form. He convincingly shows that for five Eastern European languages with complex inflection plus English, using a morphological analyzer improves performance of a tagger. He concludes that for highly inflectional languages "the use of an independent morphological dictionary is the preferred choice more annotated data". He redefines the tagging task as a choice among the tags proposed by the dictionary, using a log-linear model trained on specific ambiguity classes for individual morphological features. Hajic (2000) demonstrates convincingly that morphological disambiguation can be aided by a morphological analyzer, which, given a word without any context, gives us the set of all possible morphological tags.

The only work on Arabic tagging that uses a corpus for training and evaluation, (Diab et al., 2004), does not use a morphological analyzer. Diab et al. (2004) perform tokenization, POS tagging and base phrase chunking using a SVM based learner.

The Morphological Analysis and Disambiguation of Arabic (MADA) system is described in (Habash and Rambow, 2005). The basic approach used in MADA is inspired by the work of Hajic (2000) for tagging morphologically rich languages, which was extended to Arabic. Habash and Rambow (2005) use SVM-classifiers for individual morphological features and a simple combining scheme for choosing among competing analyses proposed by the dictionary.

3 Arabic word segmenter

Arabic is a morphologically complex language. Compared with French, an Arabic word can sometimes correspond to a whole French sentence (Example : the Arabic word "أتتذكروننا" corresponds in French to the sentence "Est-ce que vous vous souvenez de nous", in English: "Do you remember us").

The aim of a morphological analysis step is to recognize word composition and to provide specific morphological information about it. For Example : the word "يعرفون" (in French: connaîssent, in English: they know) is the result of the concatenation of the prefix "ي" indicating the present and suffix "ون" indicating the plural masculine of the verb "عرف" (in French: connaître, in English: to know). The morphological analyzer determines for each word the list of all its possible morphological features.

In Arabic language, some conjugated verbs or inflected nouns can have the same orthographic form due to absence of vowels (Example : non-voweled Arabic word "فصل" can be a verb in the past "فَصَلَ" (He dismissed), or a masculine noun "فَصْلٌ" (chapter / season), or a concatenation of the coordinating conjunction "فَ " (then) with the verb "صل": imperative of the verb (bind)).

In this work, In order to handle the morphological ambiguities, we decide to use MORPH2 (Belguith et al., 2006), an Arabic morphological analyzer developed at the Miracl laboratory[1]. MORPH2 is based on a knowledge-based computational method. It accepts as input an Arabic text, a sentence or a word. Its morphological disambiguation and analysis method is based on five steps:

- A tokenization process is applied in a first step. It consists of two sub-steps. First, the text is divided into sentences, using the system Star (Belguith et al., 2005), an Arabic text tokenizer based on contextual exploration of punctuation marks and conjunctions of coordination. The second sub-step detects the different words in each sentence.
- A morphological preprocessing step which aims to extract clitics agglutinated to the word. A filtering process is then applied to check out if the remaining word is a particle, a number, a date, or a proper noun.
- An affixal analysis is then applied to determine all possible affixes and roots. It aims to identify basic elements belonging

[1] http://www.miracl.rnu.tn

to the constitution of a word (the root and affixes i.e. prefix, infix and suffix).

- The morphological analysis step consists of determining for each word, all its possible morpho-syntactic features (i.e, part of speech, gender, number, time, person, etc.). Morpho-syntactic features detection is made up on three stages. The first stage identifies the part-of-speech of the word (i.e. verb "فعل", noun "اسم", particle "أداة" and proper noun "اسم علم"). The second stage extracts for each part-of-speech a list of its morpho-syntactic features. A filtering of these feature lists is made in the third stage.
- Vocalization and validation step : each handled word is fully vocalized according to its morpho-syntactic features determined in the previous step.

In our method, each Arabic word, from Arabic data, is replaced by its segmented form, where stem, clitic and affix are featured with their morphological classes (e.g. proclitic, prefix, stem, suffix and enclitic). For example: the word "فعرفناهم" (in French: "et nous les avons connu", in English: "and we have known them") is the result of the concatenation of the proclitic "فَ" (then): coordinating conjunction, the suffix "نا" for the present masculine plural, enclitic "هم" (for the masculine plural possession pronoun), and the rest of the word "عرف" indicating the stem. So, the word "فعرفناهم" will be replaced by:

"enclitic_ هم suffix_ نا **Stem_**عرف proclitic_ف"

4 Problems in context Morpho-syntactic feature choice

As mentioned in section 1, ambiguities in Arabic word are mainly caused by the absence of the short vowels. Thus, a word can have different meanings. There are also the usual homographs of uninflected words with/without the same pronunciation, which have different meanings and usually different POS's. For example: the word "ذهب", in English: "gold" and "ذهب", in English: "go". In Arabic there are four categories of words: noun, proper noun, verbs and particles. The absence of short vowels can cause ambiguities within the same category or across different categories. For example: the word "بعد" corresponds to many categories (table 1).

meanings of a word "بعد"	Categories
after	Particule
remoteness	Noun
remove	Verb
go away	Verb

Table 1. Different meanings of a word "بعد"

Arabic uses a diverse system of prefixes, suffixes, and pronouns that are attached to the words (Soudi, 2007).

In fact, in Arabic language, the word: "وضع" can be a verb in the past (He filed), or a masculine noun (state), or a concatenation of the coordinating conjunction "و" (and) with the verb "ضع": imperative of the verb (filed). For this reason, correct morphological analysis is required to resolve structural ambiguities among Arabic sentence.

5 Word Morpho-syntactic feature disambiguation

5.1 Training corpus

The training corpus used in this work is an Arabic-French bitext aligned at the sentence level. Each Arabic word, from Arabic data, is replaced by its segmented form. In the other side, the French corpus is part-of-speech (POS) tagged by using treetagger tool (Schmid, 1994) for annotating text with part-of-speech and lemma information.

5.2 Alignment model

The aligned model was trained with GIZA++ (Och and Ney, 2003), which implements the most typical IBM and HMM alignment models. The alignment model used consists of IBM-1, HMM, IBM-3 and IBM-4.

5.3 Using treetagger for Arabic Word Morpho-syntactic feature disambiguation

To pre-process the Arabic data, we use the MORPH2 morphological analyzer (Belguith et al., 2006). A sample output of the morphological analyzer is shown in Figure 1.

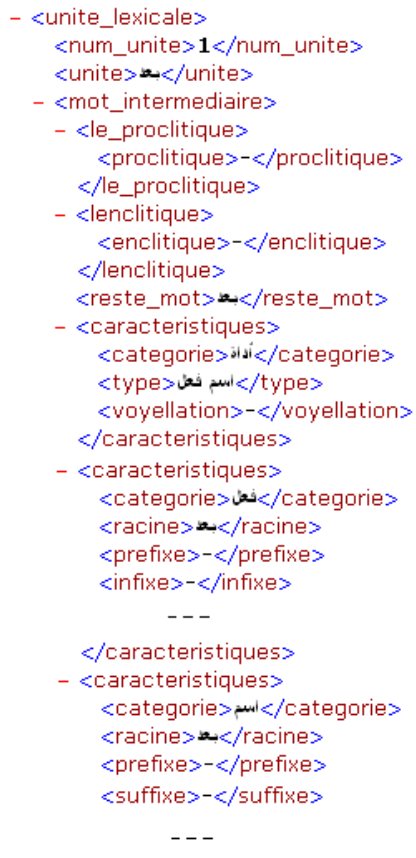

Figure 1. Possible analyses for the word "بعد"

The obtained output consists of a set of all possible morphological analyses for each word, with the unique correct analysis. One needs to select the right meaning by looking at the context. Given the highly inflection nature of Arabic, resolving ambiguities is syntactically harder within the same category. We want to choose the correct output using the features generated by TreeTagger applied to the French corpus.

To obtain this correct feature, we needed to match data in the segmented Arabic corpus to the lexeme and feature representation output by TreeTagger. The matching included the results of the alignment of word between the segmented Arabic corpus and the part-of-speech tagged French corpus.

Example : the word "فعرفناهم" (in French: "et nous les avons connu", in English: "and we have known them") is segmented by:

"enclitic_هم suffix_نا **stem**_عرف proclitic_ف"

The part-of-speech tagged of the French sentence is:

"et_KON nous_PRO:PER les_PRO:PER avons_VER:pres connu_VER:pper"

The result of the alignment between these two sentences is:

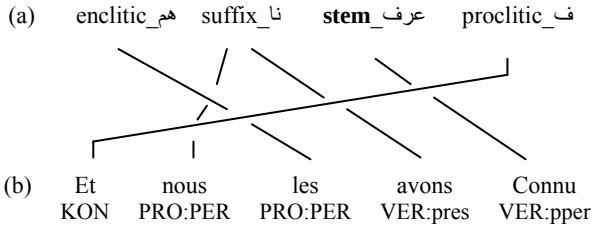

Figure 2. (a) Output of morphological analysis MORPH2: segmented Arabic sentence, (b) French translation and its alignment with segmented morphological analysis

The stem "عرف" is aligned with the word: "connu" : the past participle of the verb "connaître" (in English : "known"). We can deduce that the part of speech of the stem "عرف" is a verb.

Morpho-syntactic features provided by TreeTagger are verb, proper noun, noun, adjective, adverb, conjunction, pronoun, preposition, etc. The morpho-syntactic feature of Arabic words aligned with French words tagged by adjective or noun will be replaced by the morpho-syntactic feature : noun: "اسم". While, the morpho-syntactic feature : adverb, conjunction, or prepostion will be replaced by the Arabic morpho-syntactic feature : particle : "أداة".

We can attest that the use of morpho-syntactic features provided by the part-of-speech tagged corpus in the other side can remove disambiguation of morpho-syntactic feature of the Arabic word provided by a morphological analyser, especially for agglutinative and inflectional languages.

5.4 Experimental results

In our experiments, on the entire corpus, the MORPH2 morphological analyzer makes 1152 errors (27%). Table 2 shows the results obtained with the morphological analyzer MORPH2, BASELINE, and the results obtained with the Arabic morphology-French alignment, treetagger-to-morph2, where Arabic morphology-French alignment is used to choose a correct

morpho-syntactic feature produced by a morphological analyzer MORPH2.

System	Accuracy (%)
BASELINE	73%
treetagger-to-morph2	88%

Table 2. Results of treetagger-to-morph2 compared against BASELINE on the task of POS tagging of Arabic text

Thus one can observe that, for Arabic, the treetagger-to-morph2 outperforms the BASELINE tagger with a significant absolute difference of 15% in tagging accuracy.

The performance of treetagger-to-morph2 is better than the baseline BASELINE. The errors encountered result from confusing nouns with verbs, particles or vice versa. This is to be caused by the presence of homographs of Arabic words, which have different meanings and different POS's.

6 Conclusion

Morphological disambiguation of Arabic is a difficult task which involves, in theory, thousands of possible tags.

In this paper, we present a system which disambiguates the output of a morphological analyzer for Arabic. Arabic is a morphologically rich language, and Morphological analysis and disambiguation are crucial stages in a variety of natural language processing applications.

We first applied an Arabic word segmentation step, to improve the alignments models. So we use the Arabic morphological analyzer MORPH2. Then, we proposed to use TreeTagger tool, able to annotate text with part-of-speech and lemma information, where each word from French corpus is agglutinated to its part-of-speech (POS). The sentence alignment between Arabic and French corpus was trained with GIZA++. The core idea is to avoid morpho-syntactic ambiguity of the Arabic words obtained in the MORPH2 output by using the part of speech of corresponding aligned French word. We showed that imposing Arabic-French alignment dependent constraints on possible sequences of analyses improves the morphological disambiguation.

Future work will focus on taking advantage of our efficient technique. We are very interested to use Word Morpho-syntactic feature disambiguation to build up an efficient French-Arabic translation system.

References

Belguith L., and Chaâben N. 2006. Analyse et désambiguïsation morphologiques de textes arabes non voyellés, *Actes de la 13ème confrence sur le Traitement Automatique des Langues Naturelles*, Leuven Belgique, 493-501.

Belguith L., Baccour L. and Mourad G. 2005. Segmentation des textes arabes basée sur l'analyse contextuelle des signes de ponctuations et de certaines particules. *Actes de la 12éme Conférence annuelle sur le Traitement Automatique des Langues Naturelles,* 451-456.

Diab M., Hacioglu K., and Jurafsky D. 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.

Habash N. and Rambow O.. 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'05), Ann Arbor, Michigan, June. Association for Computational Linguistics. 573–580

Hajic J. 2000. Morphological tagging: Data vs. dictionaries. *In 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL'00)*, Seattle, WA.

Och F. J., and Ney H. 2003. A Systematic comparison of various statistical alignment models, *Computational Linguistics*, 29(1): 19-51.

Och F. J., Ney H. 2004. The alignment template approach to statistical machine translation, *Computational Linguistics*, 30(4): 417-449.

Schmid H. 1994. Probabilistic Part-of-speech Tagging Using Decision Trees. *In Proceedings of International Conference on New Methods in Language Processing*, Manchester. 44-49.

Soudi A., Bosch A. and Neumann G. Arabic Computational Morphology: Knowledge-based and Empirical Methods. Springer, 2007.