TOWARDS RESOLVING AMBIGUITY IN UNDERSTANDING ARABIC SENTENCE

Eman Othman
Computer Science Dept., Institute of Statistical Studies and Research (ISSR), Cairo Univ.
5 Tharwat St., Orman, Giza, Egypt
E-mail:emy_othman@hotmail.com

Khaled Shaalan
Computer Science Dept., Faculty of *Computers and Information, Cairo Univ.*
5 Tharwat St., Orman, Giza, Egypt
Email:shaalan@claes.sci.eg

Ahmed Rafea
Computer Science Dept., American University in Cairo
113, Sharia Kasr El-Aini, P.O. Box 2511, 11511, Cairo, Egypt.
*E-mail:*rafea@aucegypt.edu

ABSTRACT

Ambiguity is a major reason why computers do not yet understand natural language. We have made great deal strides towards developing tools for morphological and syntactic analyzers for Arabic in recent years. The absence of diacritics, which represent most vowels, in the written text creates ambiguity which hinders the development of Arabic natural language processing applications. Thus, ambiguity increases the range of possible interpretations of natural language. In this paper, we give a road map of solutions to common ambiguity problems inherent in parsing of Arabic sentence.

1. INTRODUCTION

Arabic is a strongly structured and highly derivational language (Kiraz, 2001). Understanding Arabic requires the treatment of the language constituents at all levels (Feddag, 1992): morphology, syntax, and semantics. Each component requires extensive study and exploitation of the associated linguistic characteristics (Khayat, 1996; Black, 2004).

Arabic morphology and syntax provide the ability to add a large number of affixes to each word which makes combinatorial increment of possible words (Ditters, 2001; Jaccarini, 2001). Most of the researches in Arabic NLP systems mainly concentrated on the field of morphological analysis (Farghaly; 1987; Rafea et al., 1993; Al-Shalabi et al., 1998; Beesley, 2000; Freeman, 2001; Darwish, 2002; Soudi et al., 2003; Tahir et al., 2003).

Parsing Arabic sentences is a difficult task (Othman et al., 2003). The difficulty comes from several sources: 1) the length of the sentence and the complex Arabic syntax, 2) The omission of diacritics (vowels) in written Arabic "altashkiil", 3) The free word order nature of Arabic sentence, and 4) The presence of an elliptic personal pronoun "alDamiir almustatir". Little efforts in Arabic syntactic analysis have been made in recent years (Shaalan et al., 1999; Ouersighni, 2001; Othman et al., 2003).

An important consideration, in the development of any Arabic natural language processing system, is the matter of resolving ambiguity. Ambiguity increases the range of possible interpretations of natural language. Our approach to resolve ambiguity is based on satisfying certain linguistic constraints during the course of parsing an Arabic sentence. These constraints are heuristic and specified with the grammar to ensure well-formedness. The advantages of this approach are that it is based on linguistic (morphological and syntactic) knowledge and easy to incorporate the grammar rules with heuristic constraints, which are capable of resolving ambiguity.

The rest of the paper is structured as follows. In, Section 2, we present a brief description of our Arabic syntax analyzer. Section 3 we describe our disambiguation approach. In Section 4, we give some concluding remarks and future directions.

2. SYSTEM DESCRIPTION

Syntactic analysis is a major step towards the development of most natural language processing applications. We have developed a syntactic analysis system for Arabic. Three natural language processing components are involved: a lexicon, a morphological analyzer (Rafea et al., 1993), and a syntactic parser (Othman et al., 2003). In the following, we briefly describe each of these components.

2.1 THE LEXICON

The lexicon is designed to reflect the word categories in Arabic—noun, verb, and particle— each with a different set of features. There are two types of features in the lexicon: syntactic features that resolve syntactic ambiguity and lexical features that resolve lexical ambiguity.

The default values of these features are stored in the lexicon and can be modified during the morphological analysis. The following describes the forms of the lexicon entry:

1. **Verbs:** A verb has a stem form and the following features:
   - ***Syntactic features:*** `voice,tense,transtivity, subject_gender, subject_number, object_gender,object_numbe r,end_case`
   - ***Lexical features:*** `subject_rationality,object_rati onality, infinitive`
2. **Nouns:** A noun has a stem form and the following features:
   - ***Syntactic features:*** `definition,gender,number,e nd_case, irregular_plural`
   - ***Lexical features:*** `adjectivability,category,r ationality`
3. **Particles:** A particle has a stem form and the category features.

2.2 THE MORPHOLOGICAL ANALYZER

In Rafea et al. (1993) we described a morphological analyzer for inflected Arabic words using an augmented transition network (ATN) technique. An exhaustive-search to traverse the ATN generates all the possible interpretations of an inflected Arabic word. The morphological analyzer is implemented in Prolog and integrated with the parser.

2.3 THE SYNTACTIC PARSER

Unification based grammar (UBG) formalism (Allen, 1995; Jurafsky et al., 2000) is used to write the Arabic grammar rules in the proposed chart parser. The grammar is implemented in SICStus Prolog 3.10. Each grammar rule has the form:

```
rule(LHS,RHS):-
      constraints.
```

Each constraint is used for one of three purposes: 1) To make the agreement between the left and right hand side of the grammar rule, 2) To reduce the syntactic ambiguity, and 3) To reduce the semantic ambiguity.

The grammar specifies the structure of the Arabic sentence. The Arabic sentence is generally classified as either nominal sentence or verbal sentence. In each case the sentence is either simple or compound. The difference between the simple sentence and the compound sentence is that the former does not have a complementary at the end of the sentence. We have developed a standard bottom-up chart parser for Arabic. In Othman et al. (2003) we described our Arabic chart parser. The system is implemented using SICStus Prolog on an IBM PC.

3. THE DISAMBIGUATION APPROACH

Our disambiguation approach is based on cooperation between the morphological analyzer and the parser. The morphological analyzer produces all possible interpretations of the inflected Arabic word. Disambiguation would be resolved by applying certain types of constraints that are defined with the grammar rules. Satisfying these constraints leads to a correct parse and in some cases it could resolve the ambiguity.

3.1 VERBAL SENTENCE DISAMBIGUATION ACCORDING TO THE AGREEMENT CONSTRAINTS BETWEEN THE INCHOATIVE "المبتدأ" AND ENUNCIATIVE "الخبر"

One possible grammar rule for the Arabic verbal sentence is the following:

`verbal_sentence → verb_phrase`

Where the verb phrase is defined as:

`verb_phrase→ verb`
`verb_phrase→ particle verb`

Consider the following input sentence:

الاجتماعان **حضرا** في المبنى ذاته

The meeting were attended in the same building (passive voice)

*The meeting came in the same building (active voice)

This grammar is inevitably ambiguous because we get two parse trees as shown in Fig. 1. The left parse tree shows parsing the sentence in the passive voice whereas the right parse tree shows parsing the sentence in the active voice.

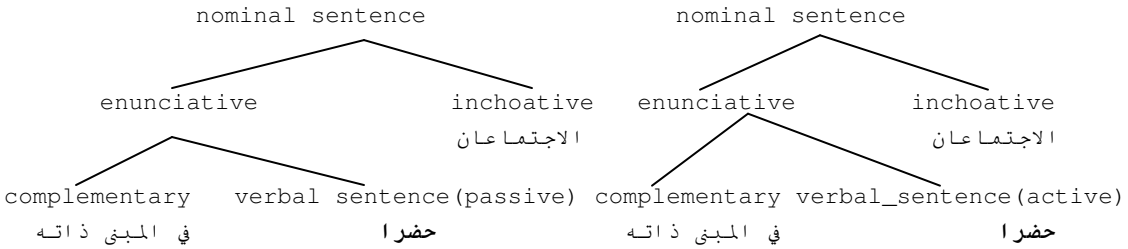

Fig. 1 Two ambiguous parse trees for the sentence الاجتماعان حضرا في المبنى ذاته

The following pseudo code explains how we recognize the possible verb voice:

Case1: intransitive verb
The verb voice is active
If there is a suffix pronoun **Then**
This pronoun is the subject
Else The subject is an elliptic personal pronoun
End if
Case2: transitive
If there is a suffix pronoun **Then**
There are two possible situations:
- The voice is active and the pronoun is the subject
- The voice is passive and the pronoun is the proagent

Else There are two possible situations:
- The voice Is active and the subject is an elliptic personal pronoun
- The voice is passive and the proagent is an elliptic personal pronoun

End if
Case3: bitransitive
The verb cannot be bitransitive in this rule
End Case

Concerning voice, there two possible parses in *case 2*. This will be resolved when we apply the agreement in rationality, gender, and number between the inchoative (مبتدأ) and enunciative (a verbal sentence) of the nominal sentence. In other words, the agreement will be applied between the inchoative and the subject features of the verb in case of active voice, and between inchoative and the object features of the verb in case of passive voice. For example, in the above sentence in Fig. 1, the inchoative is irrational, the subject rationality of the verb " حضر" is rational, and the object rationality is irrational. This leads to choose the voice of the verb as passive rather than active voice.

3.2 VERBAL SENTENCE DISAMBIGUATION ACCORDING TO THE AGREEMENT BETWEEN THE SUBJECT AND VERB

Another ambiguous grammar rule for the Arabic verbal sentence definition is as follows:

verbal_sentence → verb_phrase noun_phrase

Where the noun phrase could be either a subject, object or proagent

Consider the following input sentence:
أكل الطعام
The food was eaten (passive voice)
*the food ate (active voice)

This grammar is also inevitably ambiguous because we get two parse trees as shown in Fig. 2. The left parse tree shows parsing the sentence in the passive voice whereas the right parse tree shows parsing the sentence in the active voice.

The ambiguity in the above sentence can be resolved by applying the verb-subject and verb-proagent agreement constraints. The left tree satisfies these constraints because the verb أكل must have a rational subject and the word ألطعام is irrational. So, the left parse tree is accepted but the right tree is rejected.

Fig. 2 Two parse trees for the ambiguous sentence أكل الطعام

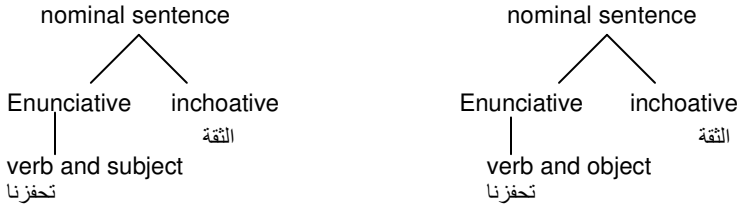

Fig. 3 Two parse trees for the ambiguous sentence الـثقـه تحفـزنـا

3.3 NOUN AFFIX DISAMBIGUATION

Ambiguous words could be dealt with during the morphological and the syntactic analysis. Consider the word "شعبي" (two nations-public level) in the following sentences

السلام كخيار استراتيجى بين شعبي فلسطين و اسرائيل

Peace as a strategic choice between palstine and Israel nations

هذين الشخصين أحدهما على مستوى حكومي والآخر على مستوى شعبي

These two persons, one of them is at the governmental level and the other is at the public level

The morphological analyzer produces all the interpretations of the word "شعبي" as follows:

1. A dual form of the noun "شعب" in the accusative case such that it is annexated to another noun
2. An adjective
3. A noun "شعب" annexated to the pronoun "ي"

Applying the grammar constraints would accept only the relevant morphological output. Accordingly, the word "شعبي" is recognized as "dual form" in the former sentence but it is recognized as "an adjective" in the later sentence.

3.4 VERB AFFIX DISAMBIGUATION

In the morphological analyzer the pronouns are classified under either nominative or accusative pronouns category. The grammar constraints use this sub-categorization to disambiguate pronouns.

Consider the following sentence:

الثقه تحفزنا لمزيد من الانجاز فى العمل

The confidence motivate us to give more achievements in the work

The pronoun 'نا' is connected to the verb 'تحفز'. There are two possible parse trees for the verbal sentence, see Fig. 3.

The left sub tree recognizes the pronoun as a subject connected to the verb and the right sub tree recognizes the pronoun as an object connected to the verb. The grammar constraints reject the left parse tree because the pronoun is an accusative pronoun and hence it cannot be parsed as a subject.

4. CONCLUSIONS

In this paper, we described our attempt to resolve certain types of ambiguity. Our approach is based on satisfying constrains among the syntactic and semantic features. We showed by examples the capabilities of our system in resolving common types of ambiguities. These concerns resolving the ambiguity that arise due to the agreement between main constituents of the nominal and the verbal sentences, and due to the affixes of either nouns or verbs. Our future work could be directed towards resolving other types of ambiguity and testing the system on large test sets.

5. REFERENCES

Allen, J. (1995). Natural Language Understanding, second edition, The Benjamin/Cummings Publishing Company.

Al-Shalabi, R. and Evens, M. (1998). A Computational Morphology System for Arabic. Workshop on Semitic Language Processing. In the Proceedings of the COLING-ACL-98, University of Montreal, Montreal, PQ, Canada, pp. 66-72.

Beesley, K. and Karttunen, L., (2000). Finite-State Non-Concatenative Morphotactics, In the Proceedings of the 38th Annual Meeting of the Association for Computational inguistics, Hong Kong, pp.191-198.

Black W., El-Kateb S., (2004). A Prototype English-Arabic Dictionary Based on WordNet. In the Proceedings of the Second International WordNet Conference— GWC 2004, Brno, Czech Republic.

Darwish K. (2002). Building a Shallow Arabic Morphological Analyzer in One Day, In the Proceedings of the Computational Approaches to Semitic Languages, A workshop affiliated with ACL-2002, University of Pennsylvania.

Ditters E. 2001. A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic, In the proceeding of Arabic NLP Workshop at ACL/EACL.

Farghaly A. (1987)., Three-level Morphology, In Proceeding of the Arabic Morphology Workshop, Stanford University.

Feddag, A. 1992. Arabic Morpho-Syntax and Semantic Parsing, in proceeding of the 3rd international conference and exhibition on Multi-lingual Computing, Univ. of Durham, UK.

Freeman A. (2001). Brill's POS Tagger and a Morphology parser for Arabic, In the proceedings of Arabic NLP Workshop at ACL/EACL

Jaccarini A. 2001. A modifiable structural editor of grammars for Arabic processing, In the proceeding of Arabic NLP Workshop at ACL/EACL.

Jurafsky D. and Martin J. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Prentice Hall.

Khayat, M. (1996). Understanding Natural Arabic, In Proceedings of the First KFUPM Workshop on Information & Computer Science, Saudi Arabia.

Kiraz G. (2001). Computational Nonlinear Morphology: with Emphasis on Semitic Languages, Cambridge University Press.

Othman E., Shaalan K., and Rafea A. (2003). A Chart Parser for Analyzing Modern Standard Arabic Sentence, In proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches, New Orleans, Louisiana, U.S.A.

Ouersighni R. (2001), A major offshoot of the DIINAR-MBC project: AraParse, a morphosyntactic analyzer for unvowelled Arabic texts, In the proceeding of Arabic NLP Workshop at ACL/EACL 2001.

Rafea A. and Shaalan K. (1993). Lexical Analysis of Inflected Arabic words using Exhaustive Search of an Augmented Transition Network, Software Practice and Experience, Vol. 23(6), pp. 567-588, John Wiley & sons, U.K.

Soudi A., Cavalli-Sforsa V. (2003). Interfacing an Arabic Morphology and Sentence Generation with an English-to-Arabic knowledge-based Machine Translation System, In the Proceedings of the workshop on Information Technology, Rabat, Morocco, March 17-19.

Shaalan K., Farouk A., Rafea, A. (1999). Towards An Arabic Parser for Modern Scientific Text, In Proceeding of the 2nd Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE), pp. 103-114, Egypt.

Tahir Y., Chenfour N., Harti M. (2003). Realization of a morphological analyzer for Arabic language text, In the Proceedings of the workshop on Information Technology, Rabat, Morocco, March 17-19.