Towards A Modern Arabic Text Machine Translation System from Arabic to English

By

Eman Othman Yousef

A Thesis Submitted to the

Department of Computer Science, Institute of Statistical Studies and Research at Cairo University

in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

in

Computer Since

Under the Supervision of

Prof. Dr. Ahmed Rafea

Computer Science Dept., Faculty of Computers and Information, Cairo University

Prof. Dr. Ibrahem Farag

Computer Science Dept., Faculty of Computers and Information, Cairo University

Prof. Dr. Khaled Shaalan

Computer Science Dept., Faculty of Computers and Information, Cairo University

٩٠٣)

Institute of Statistical Studies and Research, Cairo University

GIZA, EGYPT

2008

APPROVAL SHEET

Towards A Modern Arabic Text Machine Translation System from Arabic to English

M.S.C Thesis

By

Eman Othman Yousef

This thesis is for M.S.C Degree in Computer Science,

Department of Computer and Information Science, Institute of Statistical

Studies and Research, Cairo University, has been approved by:

Name	Signature
Prof .Dr.Mahmoud Riad	
Prof. Dr.Ibrahim Farag	
Prof. D. Ahmed Rafea	Ahmed Rafea
Prof .Dr.Ali Fahmey	

2008

Towards A Modern Arabic Text Machine Translation System from Arabic to English

By

Eman Othman Yousef

A Thesis Submitted to the

Department of Computer Science, Institute of Statistical Studies and Research at Cairo University

in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Since

Under the Supervision of

Prof. Dr. Ahmed Rafea
Computer Science Dept., Faculty of Computers and Information, Cairo University
Prof. Dr. Ibrahem Farag
Computer Science Dept., Faculty of Computers and Information, Cairo University
Prof. Dr. Khaled Shaalan
Computer Science Dept., Faculty of Computers and Information, Cairo University

Institute of Statistical Studies and Research, Cairo University
GIZA, EGYPT
2008

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation and gratitude to the supervisors’ committee Prof. Dr. Ahmad Rafea , Prof. Dr. Ibrahem Farag and Dr.Khaled Shallan for their contribution to the thesis .

I would like to express my appreciation to Prof. Dr. Ahmad Rafea for his guidance, and for enriching me with his precious and valuable knowledge and experience.

I would like to extend my gratitude to Dr. Khaled Shaalan. I am greatly indebted to him for his valuable input and his continues support and direction. I want to thank him for suggesting the research topic, encouraging me to continuously develop my research skills and for his continues encouragement. Without his generosity with his time and effort, this work would not have attained its present form.

Deep thanks to my father who teach me how to love our Arabic language and encourage me to continue my study.I want also to extend my thanks to every one in my family for the support they gave to me during the thesis preparation.

ABSTRACT

Machine translation considered as the most difficult application for NLP, a deep analysis for the source language is needed. So we can say that parsing is the backbone of any machine translation system that is a good parser is needed to support the translation process.

Language is the fundamental means of communication for human beings. Though simple and comprehensive as it may appear to human, it is in fact of most complexity when it comes to understanding in the part of the computer. Natural language processing (NLP) is the engineering of systems that process or analyze written or spoken natural language.

The aim of the current work is to develop an Arabic parser, which parse modern Arabic sentences. The parser is a chart parser. The research is targeted at parsing modern sentences as the sentences found in newspapers. The parser is being developed using Prolog language.A major design goal in developing the proposed parser is that it can be used as a tool in developing other Arabic NLP applications such as (tutoring systems, information extraction, MT system, etc.)

The parser is implemented in Prolog using unification based grammar formalism. Experiment on sentences from two different domains was performed. The thesis reports the results from the experiments and the analysis of these results.

TABLE OF CONTENTS

Chapter 1
Introduction

Machine translation considered as the most difficult application for NLP, a deep analysis for the source language is needed. So we can say that parsing is the backbone of any machine translation system that is a good parser is needed to support the translation process. This chapter will briefly sketch some background on natural language processing. Then the focus turns to a description of the scope of the current work. Next, the aim of the current work and the structure of this thesis are summarized

1.1 Natural language processing

Language is the fundamental means of communication for human beings. Though simple and comprehensive as it may appear to human, it is in fact of most complexity when it comes to understanding in the part of the computer. Natural language processing (NLP) is the engineering of systems that process or analyze written or spoken natural language. Since most of human knowledge is recorded in linguistic form, enabling computers to understand natural language would allow it to access all this knowledge (URL2).

Number of problems faces the worker in this filed such as:

- Richness: in natural languages there are many grammatical patterns and hundreds of thousands of words.
- Ambiguity: there are words of several sense and sentences of different readings.
- Creativity: New words introduced and the use of unconventional grammar.
- Indirect expression: The ideas we express are often indirectly related to the meanings of our utterances
- Implicit links among sentences: Pronoun reference, temporal reference. Implicit causal links.

(Mats D., 2004)

1.2 The standard Paradigm for Natural Language Processing

This section describes the four major aspects of natural language systems [Zarri,1998], [Zamora, 1994].

- **Lexicon**

A lexicon is a dictionary containing the words that are recognized by a natural language system. It is necessary to define what a word is before a lexicon can be constructed. The technique for isolating words affects the content of the dictionary and the applications that use the dictionary. Should the dictionary include contractions like "can't" or hyphenated words like "mother-in-low"? If capitalized words are allowed? Will the dictionary contain multiple words such as " hot dog " or abbreviations like "etc."

Having isolated the words in the input sequence, the first step of the analysis concerns the computation of their grammatical category (part of speech "tag"), i.e., the association with information like "noun", "verb" adjective" etc this particular information will then be used in all the subsequent phases of the procedure.

- **Morphology:**

The decomposition of words into their uninflected root forms, performed at the word level. There are many morphological phenomena: almost all languages have inflectional morphology; the majority has some form of derivational morphology. A number of general models of morphological processing have been investigated. At the theoretical level, the most popular approach to morphology is the so-called two-level approach (Koskenniemi 1983). In practical systems many other, less general and more language and task-specific approaches have been used.

- **Syntax:**

Syntax analysis is the extraction of all well-formed syntactic structures and dependencies for a source text, performed at the sentence level. There are many grammar formalisms, such as, for instance, Lexical Functional Grammar, Generalized Phrase Structure Grammar, Head-driven Phrase

Structure Grammar, Definite Clause Grammar, Tree-adjoining Grammar or Government-and-Binding-related Grammars. The use of a "canonical" formalism facilitates the use of a single grammar interpreter applicable to any language whose grammar is defined in the selected formalism.

- **Semantic:**

Semantic analysis is the creation of the knowledge structures in a text-meaning representation language that reflect the meanings of lexical units in the source text and semantic dependencies among them, performed at the sentence level but often having to take into account suprasentential contexts. Semantic analysis procedures are typically developed for a particular domain (e.g. medicine, finance, and computers), though general, "common sense" semantic knowledge is also used. The existence of canonical formalisms for encoding world knowledge and text meaning enables the use of a single universal semantic interpreter with different knowledge source for each domain.

1.3 Scope and Aim of the Work

The aim of the current work is to develop an Arabic parser, which parse modern Arabic sentences. The parser is a chart parser. The research is targeted at parsing modern sentences as the sentences found in newspapers. The parser is being developed using Prolog language. A major design goal of this system is that it can be used as a stand-alone tool and can be very well integrated with some other Arabic applications such as machine translation systems, systems for teaching Arabic, and system for checking and correcting grammatical errors.

1.4 Thesis Structure

This thesis is organized as follows. Chapter 2 is a review on the language under work (Arabic), a brief description of word classes in Arabic and sentences types are presented. In Chapter 3 we propose the different ways to represent language grammar and the fundamentals parsing techniques.

Chapter 4 describes the lexicon and the morphological analyzer used in the proposed Arabic parser. Chapter 5 proposes a unification based grammar for Arabic. Chapter 6 describes a proposed chart parser for Arabic. in chapter 7 we present our evaluation methodology to evaluate the Arabic Parser, the test results analyses and a discussion for that also are represented. And in chapter 8 there is the conclusion and the future work.

Chapter 2

Aspects of Arabic Language

The work in the NLP field requires a great knowledge for the language under consideration. The classification of Arabhic wrods are presented in Scetion 2.1. The classification of Arabic sentence are introduced in Section 2.2.

2.1 Arabic Words

Arabic words are classified to three main categories: Nouns, Verbs, and Particles. (الحروف).

2.1.1 Nouns

The noun is word that indicates a human, animal, plant, solid body, place, time, adjective, and and any word that its meaning does not regard to time. (فؤاد نعمة، ١٩٨٦), (Mokhtar, 2000). The nouns could be divided (subcategorized) into many divisions according to many criteria's:

- **Noun Kinds (أنواع الاسم)**

There are more than 21 kind of names, including: the Accusative (الظرف), the pronoun (الضمير), the Demonstrative (اسم الإشارة), among others. (Pierre K., 1973)

- **Conjugability (التصرف)**

a) Fully inflected (متصرف)& Diptote (غير متصرف)

The fully inflected noun is the noun that could take dual form (المثنى), plural form (الجمع), diminuation form (صيغة تصغير), as well as allowing another noun to be referred to him (ينسب أليه).

The fully inflected nouns are divided into:

-Primitive nouns (اسماء جامدة) which are not driven from any root, they include common nouns (اسماء الجنس) for example 'طفل', and proper nouns (اسماء العلم) as 'مصر'.

-Derivative nouns (اسماء مشتقة) which are derived from verbs such as Active participle (اسم فاعل) as 'عاصم', and passive participle (اسم مفعول) as 'معصوم'. (Dahdah A., 1985).

The Diptote noun has only one form regardless of its gender and number. They are(Thabet T., 1993), (Pierre K., 1973):

-Pronouns (الضمائر) such as 'انا'.

-Question nouns (اسماء الاستفهام) such as 'هل'.

-Demonstrative nouns (اسماء الإشارة) such as 'هذه'.

-Conjunctive nouns (الأسماء الموصولة) such as 'التي'.

-Noun of number (اسم عدد) such as 'خمسون'.

-Verbal noun (اسم فعل) such as 'حذار'.

-Condition nouns such as (حيثما، كيفما).

-Accusative such as 'أمس'.

b) Declinable (معرب) & Indeclinable (مبني) Noun

Declinable noun is the noun where the case of its end changes according to its position in the sentence and take one of three forms nominative (مرفوع), accusative (منصوب) or genitive (مجرور). (Dahdah A., 1985), (Pierre K., 1973)

Declinable nouns are of two kinds:

- Perfectly declinable (منصرف) nouns where the singular and broken plural (جمع التكسير) takes genitive case (يجر) with Kasrah (الكسرة) as 'كتاب'

- Imperfectly declinable (غير منصرف) noun where the singular and broken plural (جمع التكسير) takes genitive case (يجر) with Fathah (الفتحة) as 'منقار'. (فؤاد نعمة، ١٩٨٦),

Indeclinable noun is the noun where the case of its end do not changes according to its position in the sentence such as 'هي'.

▪ Structure

According to the word structure nouns are classified to:

-Abbreviated (مقصور) nouns which end with Alif of adherence (الف لازمه) such as 'عصى'.

-Prolonged (ممدود) which end with augmented Alif (الف زائدة) then Hamza such as 'سماء'.

-Defective (منقوص) noun which ends with Ya of adherence (ياء اللزوم) that is preceded with Kasrah such as 'قاضي'.

-Perfect (صحيح) noun which ends with perfect article (حرف صحيح) except the Hamza such as 'محمد'.

-Quasi_perfect (شبيه بالصحيح) noun which ends with Waw or Ya that is preceded with quiescence article (حرف ساكن) such as 'دلـــو'. (Dahdah A., 1985), (Pierre K., 1973)

▪ Meaning

-The Adjective (الصفة) and the Substantive (الموصوف): Substantive is a noun for a person, animal, thing or meaning. The Adjective is a noun annexed to the substantive to describe him/her/it.

-Indefinite (نكره) and definite (معرفة).

-Masculine (مذكر) and feminine (مؤنث).

-Singular, dual, and plural.

-Relative noun (اسم منسوب).

-Diminutive noun (اسم مصغر). (Dahdah A., 1985), (Pierre K., 1973)

2.1.2 Verbs

The verb is any word that indicates the occurrence of an action. (Mokhtar, 1997). Verbs are classified according to the following criteria:

▪ Tense (الزمن)

According to the tense verbs are divided to three main categories:

-**Past**: indicates that an action that were happened before the time of the speaking.

-**Present**: indicates that an action that were happened at the time of the speaking.

-**Imperative**: there is a demand for an action to take place in the future. (فؤاد نعمة، ١٩٨٦)