Ain Shams University
Faculty of Computer & Information Sciences
Scientific Computing Department

Abstract

of the master's thesis entitled

Computerized Translation of Arabic Noun Phrase into English

submitted as a partial fulfillment of the requirements for the degree of Master of Science in Computer and Information Sciences.

By

Ola Mohammad Ali Mohammad

B.Sc. in Computer and Information Sciences,
Demonstrator at Scientific Computing Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

Under Supervision of

Prof. Dr. Mohamed Saied Abdel wahab

Ex-Dean of the Faculty of Computer and Information Sciences,
Ain Shams University.
Professor in Scientific Computing

Prof. Dr. Mahmoud El-Said Ahmed GadAlla

Professor in Military Technical Collage

Cairo 2009

ABSTRACT

In the modern world, there is an increased need for language translation. Attempts of language translation are as old as computer themselves. The term 'machine translation' (MT) refers to computerized systems responsible for the production of translations with or without human assistance. In current Natural Language Processing (NLP) technology, however, machine translation relies heavily on expensive resources, such as large parallel corpora and expensive tools such as parsers and semantic taggers. Consequently, the number of languages that have such advanced technology at their disposal is small. While machine translation industrial technologies are mainly rule-based, current research is mainly on data-driven methods such as Statistical Machine Translation (SMT). Most SMT stems rely on parallel corpora, and the development of a Rule-Based Machine Translation (RBMT) system is a tedious and very expensive undertaking. Taggers and shallow rule-based parsers are relatively easy to obtain. Similarly many SMT approaches are hard tasks since sufficient parallel material is needed to model the whole translation process. On the other hand, more and more monolingual corpora of reasonable size are becoming available for an ever-increasing set of languages. Therefore, investigation of machine translation with limited resources is receiving more and more attention. Many researches suggested that a hybrid approach is the way to go.

The aim of this work is to build a hybrid machine translation system to translate Arabic noun phrase into English using only minimal resources for both the source and the target language. A hybrid rule-based statistical machine translation system is presented to translate Arabic Noun Phrase (NP) into English. Rule-based methods are used where representations and decisions can be determined a-priori with high accuracy based on linguistic insight. Corpora serve as a basis to ground decisions where uncertainty remains. SMT methods are used for target language generation, using only a target language corpus and a bilingual dictionary instead of a parallel corpus. 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 a general machine translation system for Arabic sentence.

Dictionary-graph based Word Sense Disambiguation (WSD) approach is used to improve machine translation using hybrid semantic-statistical method based on computing words relatedness and a statistical measure of association to get the relation between ambiguous words. This relation was used with Viterbi search algorithm to find the appropriate translation of the Arabic NP. A shallow source

language analysis, combined with a translation dictionary and a mapping system of source language phenomena into the target language and a target language corpus for generation are all the resources needed in the proposed system.

This work dealt with five statistical measures of association methods and three semantic relatedness measures. Dice method was the statistical measures of association method which gave the highest WSD accuracy of 63.8 % while Vector method was the semantic relatedness measure with the highest WSD accuracy of 53.52%. The hybrid semantic-statistical method improved the accuracy of WSD by 4.28% and dice-vector combination is the hybrid measure which gave the highest WSD accuracy of 68.08%.

The improvement in WSD strongly affects the accuracy of MT. In the baseline MT there were 170 ambiguous NPs when applying WSD using dice the number of ambiguous NPs decreased to be 39 phrases with translation accuracy of about 69%. WSD with vector decreases the number of ambiguous NPs to be 56 phrases with translation accuracy of 60%. WSD with hybrid dice-vector gave the highest improvement in translation process it decreased the number of ambiguous NPs to be 32 phrase with translation accuracy of about 73%. These results prove that hybrid method is the way to machine translation with limited resources.

Ain Shams University
Faculty of Computer & information Sciences

Computerized Translation of Arabic Noun Phrase into English

A Thesis Submitted to

Department of Scientific Computing

In partial fulfillment of the requirements for

The degree of Master of Science

by

Ola Mohammad Ali Mohammad
B.s.c in Computer & Information Sciences
(Scientific Computing)

Under Supervision of

Prof. Dr. Mohamed Saied Abdel Wahab
(Professor in department of Scientific Computing)

Prof. Dr. Mahmoud El-Said Ahmed GadAlla
(Professor in Military Technical Collage)

Cairo 2009

DEDICATION

To my Lord: "The One who created me, and guides me. The One who feeds me and quenches my thirst. And when I get sick, He cures me. The One Who puts me to death, then brings me back to life. The One Who will hopefully forgive my sins on the Day of Judgment. My Lord, grant me wisdom, and include me among the righteous."

– Prophet Abraham peace be upon him

ACKNOWLEDGMNETS

- I would like to show my full gratitude to ***Prof. Dr. Mohammad Saied Abdel wahab*** for his valuable notes through all my academic years and during thesis preparation. ***Prof. Dr. Mahmoud Gadalla*** for his guidance, constructive criticism and support during this work.
- I also extend my gratitude to ***Dr. Ahmad Farouk*** who provided me with the Arabic parser.
- I would also like to thank my friends who were encouraging and urging me continuously to finish this work.
- Finally, I would like to express sincere appreciation to my parents and my brothers who always thought that I could do it, and saw me through it to the end. I am thankful for their support and thankful to my parents for being proud of me.

ABSTRACT

In the modern world, there is an increased need for language translation. Attempts of language translation are as old as computer themselves. The term 'machine translation' (MT) refers to computerized systems responsible for the production of translations with or without human assistance. In current Natural Language Processing (NLP) technology, however, machine translation relies heavily on expensive resources, such as large parallel corpora and expensive tools such as parsers and semantic taggers. Consequently, the number of languages that have such advanced technology at their disposal is small. While machine translation industrial technologies are mainly rule-based, current research is mainly on data-driven methods such as Statistical Machine Translation (SMT). Most SMT stems rely on parallel corpora, and the development of a Rule-Based Machine Translation (RBMT) system is a tedious and very expensive undertaking. Taggers and shallow rule-based parsers are relatively easy to obtain. Similarly many SMT approaches are hard tasks since sufficient parallel material is needed to model the whole translation process. On the other hand, more and more monolingual corpora of reasonable size are becoming available for an ever-increasing set of languages. Therefore, investigation of machine translation with limited resources is receiving more and more attention. Many researches suggested that a hybrid approach is the way to go.

The aim of this work is to build a hybrid machine translation system to translate Arabic noun phrase into English using only minimal resources for both the source and the target language. A hybrid rule-based statistical machine translation system is presented to translate Arabic Noun Phrase (NP) into English. Rule-based methods are used where representations and decisions can be determined a-priori with high accuracy based on linguistic insight. Corpora serve as a basis to ground decisions where uncertainty remains. SMT methods are used for target language generation, using only a target language corpus and a bilingual dictionary instead of a parallel corpus. 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 a general machine translation system for Arabic sentence.

Dictionary-graph based Word Sense Disambiguation (WSD) approach is used to improve machine translation using hybrid semantic-statistical method based on computing words relatedness and a statistical measure of association to get the

relation between ambiguous words. This relation was used with Viterbi search algorithm to find the appropriate translation of the Arabic NP. A shallow source language analysis, combined with a translation dictionary and a mapping system of source language phenomena into the target language and a target language corpus for generation are all the resources needed in the proposed system.

This work dealt with five statistical measures of association methods and three semantic relatedness measures. Dice method was the statistical measures of association method which gave the highest WSD accuracy of 63.8 % while Vector method was the semantic relatedness measure with the highest WSD accuracy of 53.52%. The hybrid semantic-statistical method improved the accuracy of WSD by 4.28% and dice-vector combination is the hybrid measure which gave the highest WSD accuracy of 68.08%.

The improvement in WSD strongly affects the accuracy of MT. In the baseline MT there were 170 ambiguous NPs when applying WSD using dice the number of ambiguous NPs decreased to be 39 phrases with translation accuracy of about 69%. WSD with vector decreases the number of ambiguous NPs to be 56 phrases with translation accuracy of 60%. WSD with hybrid dice-vector gave the highest improvement in translation process it decreased the number of ambiguous NPs to be 32 phrase with translation accuracy of about 73%. These results prove that hybrid method is the way to machine translation with limited resources.

TABLE OF CONTENTS

ACKNOLAGEMENTS……ii
ABSTRACT……..iii
TABLE OF CONTENTS.....v
LIST OF FIGURES vii
LIST OF TABLES..... viii
LIST OF ABBREVIATIONS..... ix

CHAPTER 1 1
INTRODUCTION 1
Overview.....2
1.1 What is Machine Translation? 2
1.2 Definition of Noun Phrases.....3
1.3 Syntactical Structure 3
1.4 Divide and Conquer Approach 4
1.5 Natural Language Processing 4
1.6 The Standard Paradigm for NLP.....5
1.6.1 Source language analysis 6
1.6.2 Target language generation.....7
1.7 Arabic Language Characteristics. 8
1. 7.1 Linguistics Varieties 8
1. 7.2 Writing System 9
1. 7.3 Morphology.....10
1. 7.4 Syntax / Grammar.....11
1.8 Problems for Arabic Machine Translation.....11
1.9 Aim of the work.....12
1.10 Structure of the Thesis 13

CHAPTER 2 14
MACHINE TRANSLATION SURVEY.....14
Overview.....15
2.1 History of Machine Translation 15
2.2 Recent Approaches of Machine Translation.....18
2.2.1 Interlingua 18
2.2.2 Transfer-Based.....20
2.2.3 Example-Based 20
2.2.4 Statistical-Based.....21
2.3 Hybrid Machine Translation 23
2.4 Noun Phrase Translation.....26
2.5 Translation of languages with Limited Resources.....28
2.6 Conclusion 31

CHAPTER 3 33
WORD SENSE DISAMBIGUATION.....33
Overview.....34
3.1 Word Sense Disambiguation.....34
3.3 Approaches to Word Sense Disambiguation 39

3.3.1 Knowledge-based Approaches....40
3.3.2 Corpus-based Approaches....44
3.3.3 Hybrid Approaches....46
3.4 Cross-lingual Word Sense Disambiguation....46
3.5 Graph-based Word Sense Disambiguation....51
3.6 Tools and resources....54
3.6.1 Software tools....54
3.6.2 Sources of Information....58
3.7 Conclusion....63

CHAPTER 4....64
STRUCTURE OF THE PROPOSED TRANSLATION SYSTEM....64
Overview....65
4.1 Source Language Analysis....66
4.2 Transfer from Source Language to Target Language....74
4.3 Target Language Generation....80
4.3.1 Viterbi algorithm....81
4.3.2 Statistical Measures of Association for Ngrams(SMA)....82
4.3.3 Semantic Relatedness and Similarity....84
4.4 Conclusion....85

CHAPTER 5....87
EXPERIMENTS AND RESULTS....87
Overview....88
5.1 Evaluation Method....89
5.2 WSD using statistical measures of association(SMA)....91
5.3 WSD using semantic relatedness....94
5.4 Hybrid semantic-statistical WSD....95
5.5 Overall accuracy of the MT....96
5.6 Conclusion....98

CHAPTER 6....99
CONCLUSION AND FUTURE WORK....99
6.1 Conclusions....100
6.2 Future Work....101

REFERENCES....103
PUBLICATIONS....110
APPENDIX A....111
APPENDIX B....121

LIST OF FIGURES

Figure 1.1 Five levels of syntactic structure.....3
Figure 1.2 Divide and conquer 4
Figure 1.3 Arabic Dialects and Sociolects.....9
Figure 1.4 Morphological structure for *taskuniyna* (you(f, sg) live).....11
Figure 2.1 The Components of Interlingual System.....19
Figure 2.2 The Components of Transfer-Based System 20
Figure 2.3 An example of connectionist association model with one MLP per output state.....26
Figure 2.4 Design of the noun phrase translation subsystem: The base model generates an n-best list that is rescored using additional features .29
Figure3.1 Example of Constructed Graph.....51
Figure 3.2. Sample graph built on the set of possible labels 52
Figure 3.3 an example of an undirected graph. 53
Figure 4.1 The architecture of the Arabic to English MT system.....65
Figure 4.2 Examples of the lexicon entries 77
Figure 4.3 The morphology analysis of the word "كمحاصيلها" 79
Figure 4.4 The parser architecture.....70
Figure 4.5 Arabic grammar constituents 70
Figure4.6 An example of structural transfer of annexation constituent. ..46
Figure4.7 An example of structural transfer of adjective constituent. 47
Figure4.8 aA example of structural transfer of substitution constituent. .47
Figure4.9 An example of NP synthesis according to the definition feature. 80
Figure4.10 An example of NP synthesis according to the number feature. 80
Figure4.11 Viterbi algorithm for finding optimal sequence of senses. 81
Figure5.1 Chart of statistical measures of association with 53649 words corpus.....91
Figure5.2 Chart of statistical measures of association with corpus of 138205 words.....92
Figure5.3 Chart of Statistical measures of association with 1171868 words corpus.....93
Figure 5.4 Chart of statistical measures of association methods with respect to corpus size. 94
Figure 5.5: Chart of semantic relatedness measures.....95
Figure 5.6: Chart of hybrid semantic-statistical measures 96
Figure 5.7 Chart of statistical, semantic and hybrid methods 97
Figure 5.8 Chart of accuracy and ambiguity of statistical, semantic and hybrid methods.....98

LIST OF TABLES

Table 3.1 How to disambiguate *interest* using a second-language corpus . 48
Table 3.2 NLTK Modules and Corpora..... 58
Table 3.3 Penn Treebank Tag set 60
Table 3.4 Noun relations in WordNet 61
Table 3.5 Verb relations in WordNet..... 61
Table 3.6 Adjective and adverb relations in WordNet..... 61
Table 3.7 WordNet entry for the noun "car" 62
Table 3.7 Hyponymy chain for the noun "*valley*" 62
Table 4.1 Examples of stem words and stem types..... 66
Table 4.2 Examples of words and the abilities to adjective 67
Table 4.3 Examples of words and the abilities to annexation..... 67
Table 4.4 Prefixes of Arabic word..... 68
Table 4.5 Suffixes of Arabic word 69
Table 4.6 Grammar of inchoative constituent and enunciative constituent 71
Table 4.7 Patterns of adjective constituent..... 72
Table 4.8 Patterns of substitution constituent..... 72
Table 4.9 Patterns of annexation constituent 73
Table 4.10 Patterns of distinguish constituent 74
Table 4.11 Patterns of digit constituent 75
Table 4.12 Examples of the Arabic-English dictionary entries 75
Table 4.13 The effect of the morphological features on the translated word 76
Table 4.14 Adjective constituent transfer rules 77
Table 4.15 Annexation constituent transfer rules..... 78
Table 4.16 Substitution constituent transfer rules..... 79
Table 4.17 Distinguish constituent transfer rules 79
Table 4.18 Digit constituent transfer rules 79
Table 4.19 Contingency table..... 83
Table 5.1 Results of Statistical measures of association with 53649 words corpus 90
Table 5.2 Results of statistical measures of association with corpus of 138205 words 91
Table 5.3 Results of Statistical measures of association with 1171868 words corpus 92
Table 5.4 Accuracy of statistical measures of association methods with respect to corpus size. 93
Table 5.5 Results of semantic relatedness measures..... 94
Table 5.6 Results of hybrid semantic-statistical measures..... 95
Table 5.7 Accuracy and ambiguity of statistical, semantic and hybrid methods 97

LIST OF ABBREVIATIONS

Abbreviations	Expanded Form
AI	Artificial Intelligent
AWN	Arabic WordNet
BSP	Bigram Statistics Package
CAT	Computer-Aided Translation
CBAG	Case-Based Analysis and Generation Module
CLIR	Cross Language Information Retrieval
DCG	Definite Clause Grammar
EBLMT	Example-Based Lexicalist Machine Translation
EBMT	Example-Based Machine Translation
HAMT	Human-Aided Machine Translation
HSO	The Hirst St-Onge measure
IR	Information Retrieval
KPP	Key Player Problem
LTM	Lexeme-Based Translation Memory
MAHT	Machine-Aided Human Translation
MI	Mutual Information
ML	Machine Learning
MRD	Machine Readable Dictionaries
MSA	Modern Standard Arabic
MT	Machine Translation
NLP	Natural Language Processing
NLTK	Natural Language ToolKit
NP	Noun phrase
NSP	Ngram Statistics Package
POS	Parts of Speech
RBMT	Rule-Based Machine Translation
SL	Source Language
SMA	Statistical Measures of Association
SMT	Statistical Machine Translation
STM	String-Based Translation Memory
TL	Target Language
TMs	Translation Memories
TWS	Target Word Selection
WSD	Word Sense Disambiguation

CHAPTER 1

INTRODUCTION

Overview

Why should we be interested in using computers for translation at all? The first and probably most important reason is that there is just too much that needs to be translated and those human translators cannot cope. A second reason is that on the whole technical materials are too boring for human translators, they do not like translating them, and so they look for help from computers. Thirdly, as far as large corporations are concerned, there is the major requirement that terminology is used consistently; they want terms to be translated in the same way every time. Computers are consistent, but human translators tend to seek variety; they do not like to repeat the same translation and this is no good for technical translation. A fourth reason is that the use of computer-based translation tools can increase the volume and speed of translation throughput, and companies and organizations like to have translations immediately, the next day, even the same day. The fifth reason is that top quality human translation is not always needed. Because computers do not produce good translations, some people do not think that they are any use at all. The fact is that there are many different circumstances in which top quality is not essential, and in these cases, automatic translation can and is being used widely. Lastly, companies want to reduce translation costs and on the whole with machine translation and translation tools they can achieve them.

Any one of these reasons alone can be sufficient justification for using and installing either MT systems or computer translation aids.

This chapter will briefly sketch some background on natural language processing and machine translation. Arabic language characteristics and the problems for Arabic machine translation are also discussed. Then the aim of the current work and the structure of this thesis are summarized.

1.1 What is Machine Translation?

In a simple description, Machine Translation (MT) is the use of computer software to translate text from one natural language into another [1]. This definition accounts for the grammatical structure of each language and uses rules and assumptions to transfer the grammatical structure of the source language (text to be translated) into the target language (translated text).

Translation is not a simple task. It is not a mere substitution for each word in a sentence with its equivalent. The process needs the faculty of being able to know 'all

of the words' in a given sentence or phrase and how one word may influence the other. Human languages consist of morphology (the way words are built up from small meaning-bearing units), syntax (sentence structure), semantics (meaning), and countless ambiguities.

1.2 Definition of Noun Phrases

The Noun Phrases (NP) of sentence are the maximal syntactic phrases that contain at least one noun and no verb. The formal definition of NP was given in [2] as:
Given a sentence *s* and its syntactic parse tree *t*, the NP/PP of the sentence *s* are the subtrees t_i that contain at least one noun and no verb, and are not part of larger sub tree that contains no verb.

1.3 Syntactical Structure

The levels of syntactic structure as illustrated in Figure 1.1 are word, base noun phrase, noun phrase (the focus of his work), clause, and discourse.

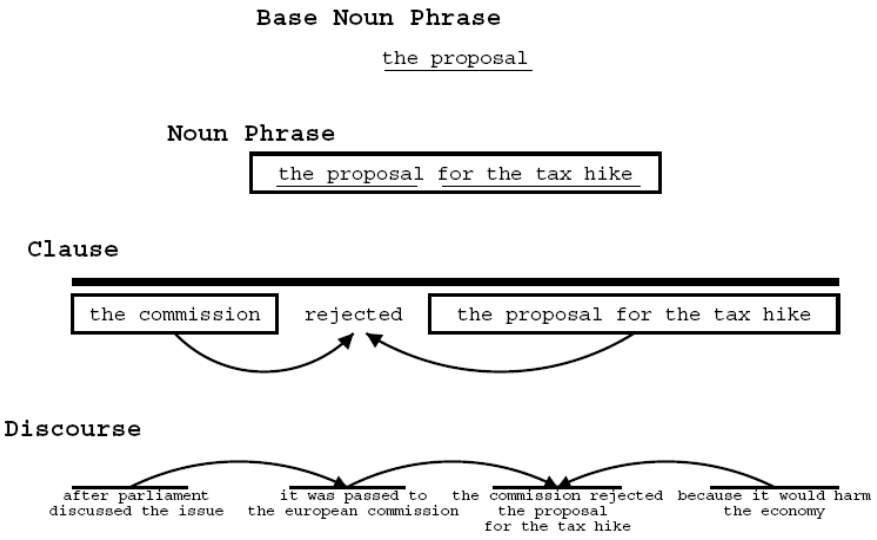

Figure 1.1 Five levels of syntactic structure

Each of these levels poses challenges for translation. Different languages may differ in their syntactic structure in general: for instance the placement of the verb in clause structure or the use of prepositions or morphology to make the role of base noun phrases. But also specific words and idiomatic expressions may force changes in syntactic structure. Ultimately, a machine translation system has to take syntactic structure into account.