Ain Shams University
Faculty of Al-Alsun (Languages)
Department of English

Pronominal Anaphora Resolution in Arabic/English Machine Translation Systems Using *Al-Ahram* Newspaper 1998-2006 as an Input

By
Rania Mostafa Mohammed Ezz Al-Sabbagh
Faculty of Al-Alsun (Languages)
Ain Shams University

Submitted in the fulfillment of the requirements for the degree of
Master of Arts

Under the Supervision of
Prof. Nagwa Taha El-Zeiny
Faculty of Arts
Helwan University

Dr. Khaled Elghamry
Faculty of Al-Alsun (Languages)
Ain Shams University

2008

To my family

Acknowledgements

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I wish to express my sincere thanks and appreciation to Prof. Nagwa El-Zeiny for her expertise, comments, stimulating feedback and support during this research and the preparation of the thesis. Without her guidance, support and good nature, I would never have been able to pursue my academic career.

I am deeply indebted to Dr. Khaled Elghamry. With his enthusiasm, his inspiration, and his great efforts to explain things clearly and simply, he helped to make Computational Linguistics fun for me. Throughout my thesis-writing period, he provided encouragement, sound advice, good teaching, and lots of good ideas.

Finally, I would like to thank my family. Throughout all my endeavors, their never-ending love, support, guidance, and patience have been truly inspirational—“thanks” will never suffice.

Abstract

Anaphora Resolution (AR) is the process of determining the antecedent of a given anaphor. It is an understudied issue in Arabic Natural Language Processing (ANLP), although some current Machine Translation (MT) systems handle it poorly. AR is usually difficult because it requires various types of knowledge and resources – syntactic, lexical and morphological – which are not available for such a language like Arabic given its scarce Natural Language Processing (NLP) resources and tools. Consequently, the proposed algorithm follows a statistical, corpus-based approach, using the Web as corpus to overcome the sparseness of data and to provide necessary resources for Arabic AR such as semantic features, collocational associations and non-pleonastic pronouns. Evaluated against a gold standard set of manually annotated pronouns, the algorithm achieves an F-measured performance rate of **87.6%**.

Table of Contents

Transliteration List

The Buckwalter Arabic Transliteration Scheme was developed at Xerox by Tim Buckwalter in the 1990s. It is an American Standard Code for Information Interchange (ASCII) transliteration scheme, representing Arabic orthography strictly one-to-one, as opposed to transcription schemes of Arabic that are intended to convey phonological or morpho-phonological information not expressed in the Arabic script. ASCII codes represent texts in computers, communication equipment and other devices that work with texts[1].

Transliteration Symbol	Arabic Letter
'	ء
>	أ
<	إ
&	ؤ
}	ئ
A	ا
B	ب
P	ة
T	ت
V	ث
J	ج
H	ح
X	خ
D	د
*	ذ
R	ر
Z	ز
S	س
$	ش
S	ص

[1] Refer to the Website of www.xrce.xerox.com

D	ض
T	ط
Z	ظ
E	ع
G	غ
F	ف
Q	ق
K	ك
L	ل
M	م
N	ن
H	هـ
W	و
Y	ى / ي

List of Tables

List of Diagrams

List of Acronyms

Acronym	Long Form
ACE	Automatic Context Extraction
AI	Artificial Intelligence
AR	Anaphora Resolution
ANLP	Arabic Natural Language Processing
ATB	Arabic TreeBank
BNC	British National Corpus
CL	Computational Linguistics
IR	Information Retrieval
ISI	Information Science Institute
MT	Machine Translation
NLP	Natural Language Processing
OOV	Out-of-Vocabulary
POS	Part-of-Speech
QA	Question Answering
SFA	Semantic Features Acquisition
SMT	Statistical Machine Translation
SNLP	Statistical Natural Language Processing

SP	Speech Recognition
SVM	Support Vector Machine
TS	Text Summarization
WSD	Word Sense Disambiguation

Part One

Theoretical Framework

1.1 Introduction

Part one contains two chapters that delineate the theoretical background of the present thesis. The first chapter introduces and defines the field of Natural Language Processing (NLP) and distinguishes it from two relevant fields, namely Computational Linguistics (CL) and Artificial Intelligence (AI). It also outlines the main criteria according to which NLP approaches are classified into different categories: supervised vs. unsupervised, symbolic vs. statistical ... etc. Therefore, the first chapter sheds light on the field of NLP in general.

The second chapter elaborates more specifically on Statistical Natural Language Processing (SNLP) which is the approach followed in the present thesis. The chapter starts with defining SNLP and highlighting the motivations for such an approach. Afterwards, it focuses on the key methods of SNLP – modeling, estimation, corpora and evaluation. Furthermore, as a part of a general introduction to the field of SNLP, the chapter discusses the problem of sparse data and its possible solutions in the framework of SNLP. Finally, some SNLP applications are pointed out.

Both chapters are an introduction to NLP in general and SNLP in particular. These chapters are to set the theoretical background of the present thesis.

Chapter One

Introduction

1.1 Introduction

This chapter briefly introduces the field of Natural Language Processing (NLP). It first starts with defining NLP and exposing the relation between NLP, Computational Linguistics (CL) and Artificial Intelligence (AI). Second, it discusses some of the most common NLP approaches and the criteria according to which these approaches are classified in order to pave the way to a more elaborated discussion of the approach of Statistical Natural Language Processing (SNLP), being the one used for the present thesis.

1.2 What is Natural Language Processing (NLP)?

Natural Language Processing (NLP) is "an area of research and application that explores how computers can be used to understand and manipulate written or spoken natural languages to perform such applications as Machine Translation (MT), Text Summarization (TS), Question Answering (QA), Information Retrieval (IR), Speech Recognition (SR) ... etc" (Chowdhury 2003: 51).

Achieving "human-like language processing" using "naturally-occurring texts" is the NLP ultimate objective according to Liddy (2001: 2126) that is usually realized using techniques from different fields of study such as formal linguistics, computer science, statistics … etc. Therefore, NLP is also defined as:

> An interdisciplinary area based on many fields of study …
> computer science, which provides techniques for model

> representation, algorithm design and implementation; linguistics, which identifies linguistic models and processes ... psychology, which studies models and theories of human behavior ... statistics, which provides techniques for predicting events based on sample data ... (Manaris 1998: 5).

The term NLP frequently collocates with Computational Linguistics (CL) and Artificial Intelligence (AI). CL is a subsection of NLP since it provides the models for the different linguistic phenomena that are used in NLP tasks. In other words, CL provides the working component of NLP systems (Ivanova 2000, Richter 2006). AI, on the contrary, is the umbrella field of which NLP is a subsection (Siekmann 2007). AI is defined as "a field of endeavor where computers and software programs are designed to mimic human reasoning and learning processes through the discovery of algorithms" (*Glossary*, 2008: 116). AI aims at developing machines, whose intelligence (i.e. their ability to process, understand and generate) mimics human intelligence. NLP is the subfield of AI that deals with designing algorithms for computers to process, understand and generate *language* in the way humans do (Elhaddad 2006).

Therefore, Natural Language Processing (NLP), Artificial Intelligence (AI) and Computational Linguistics (CL) are closely related through common objectives and techniques. The relation between the three fields can be summarized in the following diagram:

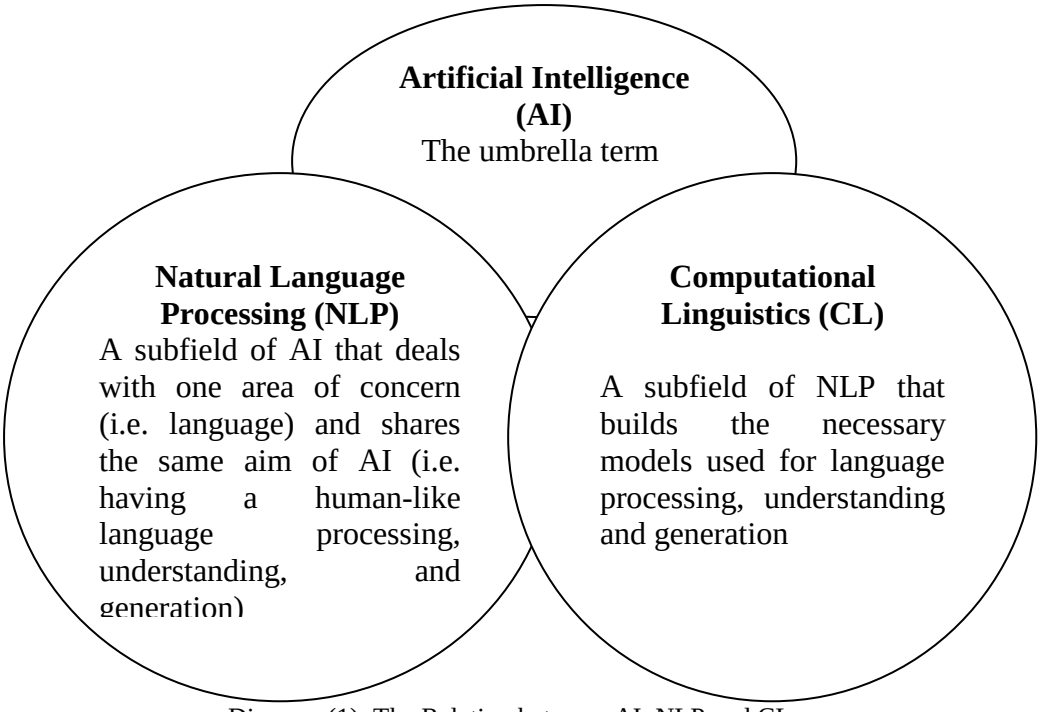

Diagram (1): The Relation between AI, NLP and CL

1.3 Approaches to NLP

As indicated by Jurafsky and Martin (2000), NLP approaches can be divided according to the following four criteria:

(1) The nature of the training material

(2) The degree of human intervention

(3) The kind of knowledge used

(4) The amount of needed material

Each one of these criteria is further subdivided into different classes as discussed in the following subsections.

1.3.1 The Nature of the Training Material

This criterion deals with the type of corpora being used. In NLP, the simplest definition of a corpus is "collections of texts and/or speech" (Burch and Osborne 2003: 5). There are different types of corpora classified according to the number of

languages being dealt with and the level of analysis involved in the way illustrated in diagram (2) designed by the researcher below:

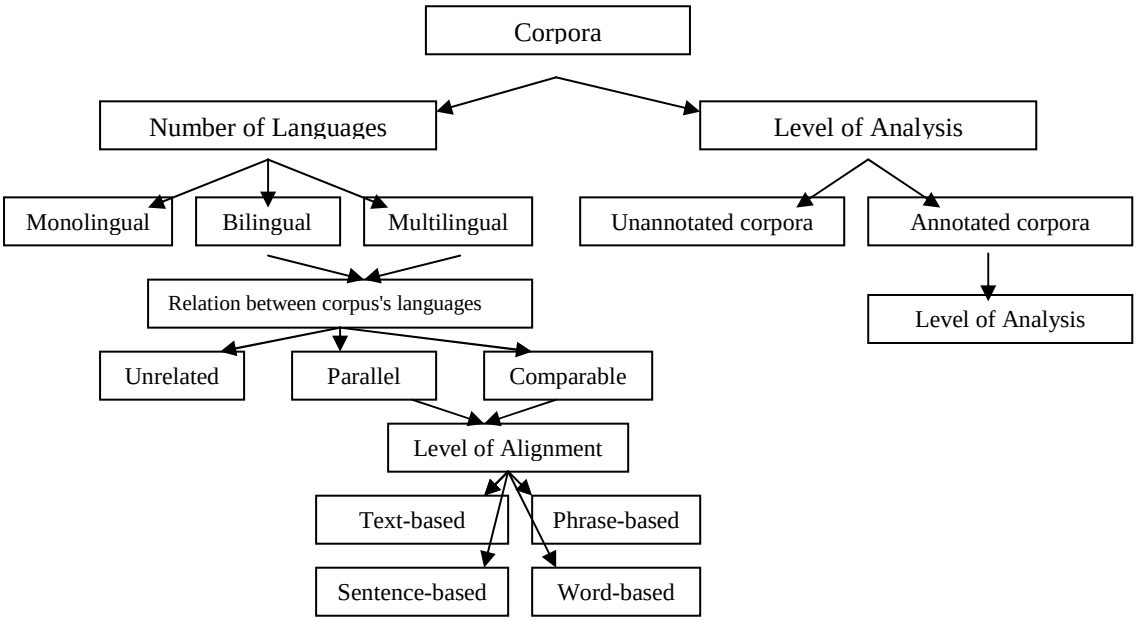

Diagram (2): Different Classifications of Corpora

According to diagram (2), corpora are divided according to the number of languages involved into:

1. Monolingual corpora which deal with one language only, such as *Al-Hayat* corpus[2] – a monolingual Arabic corpus – and the British National Corpus (BNC)[3] – a monolingual English corpus.

2. Bilingual corpora which include two languages only like the Arabic/English Parallel News Text[4].

[2] The corpus contains *Al-Hayat* newspaper articles that are distributed into 7 subject-specific databases. Mark-up, numbers, special characters and punctuation marks are removed. The size of the total file is 268 MB. The dataset contains 18,639,264 distinct tokens in 42,591 articles, organized in 7 domains. The reader is referred to http://www.elda.org/catalogue/en/text/W0030.html for more information.

[3] The British National Corpus (BNC) is a 100-million-word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written. The latest edition is the *BNC XML Edition*, released in 2007. For more information, the reader is referred to http://www.natcorp.ox.ac.uk/corpus/index.xml

[4] This corpus contains Arabic news stories and their English translations collected via Ummah Press Service from January 2001 to September 2004. It totals 8,439 story pairs, 68,685 sentence pairs, 2M Arabic words and 2.5M English words. The corpus is aligned at sentence level. It is available through Linguistic Data Consortium (LDC) catalog number LDC2004T18, URL: http://www.ldc.upenn.edu/

3. Multilingual corpora which include at least three languages such as the Automatic Context Extraction (ACE) 2004 Multilingual Training Corpus[5].

Bilingual and multilingual corpora are further subdivided according to the relation between the languages involved. On the one hand, a bilingual/multilingual corpus might consist of collections of unrelated, completely different monolingual corpora like the previously mentioned ACE 2004 Multilingual Training Corpus (McEnery and Wilson 2001). On the other hand, languages in a bilingual/multilingual corpus might be related.

In the case that the languages within a bilingual/multilingual corpus are related, the corpus is either parallel or comparable. According to McEnery and Xiao (2005), a parallel corpus holds 'the same' texts in more than one language; that is, a parallel corpus consists of texts existing in translation in two different languages, primarily translated by hand such as the aforementioned Arabic/English Parallel News Text. A comparable corpus – such as Information Science Institute (ISI) Arabic-English Automatically Extracted Parallel Text[6] – is a collection of 'similar' texts in different languages or in different varieties of a language combined on the basis of similarity of content, domain and/or communicative function.

Whether the corpus is parallel or comparable, it is usually aligned at a certain level. Alignment refers to noting which texts, paragraphs, sentences, phrases or words

[5] This corpus contains the English, Arabic and Chinese training data for the 2004 Automatic Content Extraction (ACE) technology evaluation. The corpus consists of data of various types annotated for entities and relations and was created by LDC with support from the ACE Program, which is a program for automatic context extraction It is available through Linguistic Data Consortium (LDC) catalog number LDC2005T09, URL: http://www.ldc.upenn.edu/

[6] This is a sentence-aligned Arabic-English comparable corpus, which is automatically extracted from news articles published by Xinhua News Agency and Agence France Presse. It is obtained using the automatic parallel sentence identification method described in Stefan, D. and Marcu, M. (2005). Machine Translation Performance by Exploiting Non-parallel Corpora, *Computational Linguistics*, Vol. 31. pp. 477-504. The corpus contains 1,124,609 sentence pairs; the word count on the English side is approximately 31M words.

in one language correspond to which texts, paragraphs, sentences, phrases or words in another language (Manning and Schütze 2002).

The level of alignment is another element that divides bilingual/multilingual corpora into different categories. Bilingual/multilingual texts can be aligned at the text level, sentence level, phrase level or the word level. As for text-aligned corpora, they can be automatically compiled in large amounts from the World Wide Web (WWW) (Resnik 1999). Sentence-aligned corpora, like Arabic/English Parallel News Text and ISI Arabic-English Automatically Extracted Parallel Text, are also available – yet they are rather rare, especially for languages such as Arabic. Finally, phrase- and word-aligned corpora are rarely available for Arabic and English languages.

According to diagram (2), the second criterion for corpora division is the level of analysis of the language(s) involved. This criterion divides corpora into annotated and unannotated corpora. *Unannotated corpora* refer to "large collections of raw text" (Burch and Osborne 2003: 5), such as *Al-Ahram Newspaper* corpus used for the present thesis. Annotated corpora "add additional information to the text such as phonetic transcription, Part-Of-Speech (POS) tagging, parse trees, rhetorical relations … etc" (Burch and Osborne 2003: 5). The level of annotation further subdivides corpora, since some corpora contain only one piece of linguistic information such as POS or parse trees; and other corpora involve more than one type of linguistic information such as the Arabic TreeBank[7] (ATB).

In brief, there are various corpora types divided according to different criteria. The number of languages involved in a corpus divides corpora into monolingual, bilingual and multilingual corpora. Whenever the languages within a bilingual or a

[7] Arabic TreeBank is a 1-million-word corpus which contains POS tags and parses. It s available through LDC, Catalogue Number LDC2005T20, URL: http://www.ldc.upenn.edu/

multilingual are related, the corpus might be parallel or comparable. In either case, the corpus must be aligned at the word, phrase, sentence or at least the text level. The level of analysis is another criterion that divides the corpora into annotated and unannotated corpora. The type of the corpus is the first criterion for determining an NLP approach.

1.3.2 The Degree of Human Intervention

Human intervention refers to the amount of human control over the input and/or the output of the algorithm. Human intervention divides NLP approaches into three categories: supervised, semi-supervised and unsupervised. Jones (2005) argues that in supervised approaches the algorithm is provided with a label for every example, and uses this information to learn a mapping from examples to labels for new instances. Formally speaking, Chapelle et al. (2006) show that supervised approaches involve estimating some function $f=X \rightarrow Y$ given a set of labeled training examples $\{\langle x_i; y_i \rangle\}$. Supervised learning is usually regarded as a classification task, where the output is either a class label or a decision rule that applies to new examples. Since input labels are usually provided via humans, supervised approaches are rather computationally expensive. Consequently, more interest is currently given to unsupervised ones.

In unsupervised approaches, no labels are provided at all. Instead, the algorithm sorts the data into related clusters, based on measures of proximity on the example features (Jones 2005). According to Chapelle et al. (2006), unsupervised approaches aim at either data clustering or feature extraction. On one hand, data clustering aims to unravel the structure of the provided data set. Feature extraction, on the other hand, often seeks to reduce the dimensionality of the data so as to provide a more compact representation of the data set. Unsupervised approaches are data-

driven, and thus they depend on the availability of corpora, which are not necessarily available (see section 2.4.3).

Semi-supervised approaches are halfway between the supervised and the unsupervised ones. Semi-supervised approaches are motivated by the unavailability of high-quality data. In addition to unlabeled data, a semi-supervised algorithm is provided with some supervision information – but not necessarily for all examples (Chapelle et al. 2006). The algorithm iteratively uses those labels and the data to learn approximate models, which are used to re-label and relearn better models (Jones 2005). Semi-supervised approaches carry the advantages of both supervised and unsupervised approaches, since they combine the fine-grained results of human intervention and save time as the input is only partially labeled.

Thus human intervention divides NLP approaches into supervised approaches, unsupervised approaches – which are the two extremes – and semi-supervised approaches – which are the halfway between the two extremes. Each approach has its own advantages and disadvantages: the unsupervised saves time and effort, yet its performance is not as fine-grained as the supervised one; the semi-supervised approach tries to combine the advantages of both.

1.3.3 The Kind of Knowledge Used

This criterion divides NLP approaches into symbolic, hybrid and statistical. S*ymbolic approaches* use human introspective knowledge – represented via rules – to model language behavior which is not clearly represented in the available data. In contrast, *statistical* approaches use observed data as their primary source of evidence to approximate generalized models of the linguistic phenomena under study based only on actual examples (Liddy 2001). As for the output, statistical models usually

result in a model, whereas symbolic approaches result in a set of rules formalizing the language studied.

In spite of the fine-grained output of symbolic approaches, interest in statistical approaches is increasing due to a number of reasons, such as the need to deal with broad and real-world contexts and the robustness of such approaches (i.e. their ability to deal with unexpected and/or noisy types of input provided that the training data is sufficient) (Liddy 2001)[8]. Moreover, NLP research is moving to *hybrid approaches* that utilize the strengths of each approach in order to address NLP problems more effectively and flexibly. Such approaches typically integrate more sophisticated linguistic information with statistical models to perform better (Burch and Osborne 2003, Kilgarriff and Grefenstette 2003, Keller et al. 2003, among others).

In brief, the knowledge involved in an NLP approach can be either symbolic – relying on human-crafted rules, statistical – relying on probabilities – or hybrid – combining both types of knowledge.

1.3.4 The Amount of Needed Material

This criterion largely depends on the previous one (i.e. the kind of knowledge used). According to Liddy (2001), statistical approaches typically require more data than symbolic approaches because they depend on actual examples extracted from the corpus, whereas symbolic approaches are fed with human knowledge. Depending on large corpora causes various problems for statistical approaches[9].

[8] See section (2.3) below for more details.
[9] See section (2.5) below for details

1.4 Summary

NLP is a research area concerned with processing, understanding and generating natural languages through computers. It is a subfield of AI and it uses the techniques of CL to build its models. There are different criteria according to which NLP approaches are divided: the nature of the training material, the degree of human intervention, the kind of knowledge used and the amount of needed material. Each of these criteria is further subdivided into different classes.

Chapter Two

Statistical Natural Language Processing

2.1 Introduction

This chapter focuses on Statistical Natural Language Processing (SNLP), being the approach used for the present thesis. It starts with defining SNLP and briefly illustrates the motivations behind its development. Afterwards, a quick review of SNLP key methods, including modeling, estimation, corpora and evaluation, is provided. Moreover, one of the main problems of SNLP, namely sparseness of data, is pointed out and light is shed on some SNLP applications that might be relevant to the present thesis.

2.2 What is Statistical Natural Language Processing (SNLP)?

Statistical Natural Language Processing (SNLP) is an NLP approach (see section 1.3.3), which relies on available data as its primary source of evidence to approximate generalized models of the linguistic phenomena under study based only on actual examples (Liddy 2001).

2.3 Motivations for SNLP

According to Abney (2000) and Burch and Osborne (2003), SNLP systems are useful for a number of reasons. First, they can handle a wide variety of input and this is required for real-world applications. Second, many SNLP systems are language independent such as Pantel and Lin (2001), who developed a statistical language-independent term extractor system and Aone and Mckee (1993), who developed a statistical language-independent AR algorithm. Such language-independent systems save both time and effort.

Third, SNLP systems are cheaper and faster than hand-crafted systems, since they are (semi-) automatically created. Besides, they are robust; that is, they can always produce some output no matter how badly-formed the input is or how novel it is. Finally, they require less knowledge of the particular language being analyzed and

thus work well with languages of scarce resources and allow many different approaches to be tested and evaluated in a short time. As a result, much interest is currently given to SNLP.

2.4 Methods of SNLP

There are four core methods in SNLP according to Burch and Osborne (2003): modeling, estimation, corpora and evaluation.

2.4.1 Modeling

Modeling is "the task of constructing machinery which mimics some task" (Burch and Osborne 2003: 7). It leads to language models which are defined as "lists of weighted words or combination of words that describe how words are related, how they are used with each other and how common they are in a given domain" (Kilgarriff and Grefenstette 2003: 11). Simply, statistical language models try to predict the next word(s) given the previous one(s) (Manning and Schütze 2002).

According to Burch and Osborne (2003), any language model typically consists of:

(1) Structural components (i.e. words, sounds, phrases, sentences, rules, idiomatic expressions … etc.)

(2) Parameters (i.e. the variables of such structural components)

(3) Instantiation (i.e. the probabilities assigned to each parameter).

The following lines briefly review each component of a language model.

2.4.1.1 Structural Components of Statistical Models

Structural components of a language model are determined according to the task or the application under study. For instance, verbs and their predicates are important for studies concerned with verb subcategorization frames (Korhonen 2002), phrases and idiomatic expressions are important structural components for

computational lexicography (Al-Hafez et al. 1994) and for Example-Based Machine Translation (EBMT) (Hutchins 2005); Prepositional Phrases (PPs) are used for studies focusing on PP attachment problems (Olteanu and Moldovan 2005) and so on.

The following subsection focuses on "collocations" as one of the widely used structural components for a variety of SNLP tasks and applications and as the structural component used for the present thesis.

2.4.1.1.1 Collocations

Different definitions of collocations are given within the SNLP framework; each definition adds a new criterion as revealed in the following lines.

A collocation is defined by Myka and Doucet (2005: 194) as "a recurrent combination of words that co-occur more often than chance and that correspond to arbitrary word usages". This definition focuses on the "arbitrary" usage of collocations which underlines the fact that if one word of a collocation is substituted by a synonym, the resulting phrase may become peculiar or even incorrect. For instance, in English "*strong tea*" cannot be replaced with "*powerful tea*", although *strong* and *powerful* are synonymous. Likewise, in Arabic, كثيف /*kvyf*/ (deep) in ضباب كثيف /*DbAb kvyf*/ (deep fog) cannot be replaced with غزير /*gzyr*/ (heavy), which is more likely to collocate with مطر /*mTr*/ (rain) as in مطر غزير /*mTr gzyr*/ (heavy rain) (Heliel 1989 as cited in Brashi 2005).

Another definition of collocations, provided by Jurafsky and Martin (2000: 637), is "a quantifiable position-specific relationship between two lexical items". Their definition focuses on the "position" of the two collocating words; which, according to them, is fixed. Furthermore their definition implies that there is a difference between collocations and word co-occurrences; the former has a fixed form, whereas the latter deals with words frequently used together regardless of their

positions (Jurafsky and Martin 2000). In other words, Jurafsky and Martin's definition (2000) focuses on adjacency as a distinctive feature of collocations.

However, for Manning and Schütze (2002), collocations do not necessarily consist of adjacent words. For some SNLP applications two words can be a collocation even if they are not consecutive. Thus the verb "knock" and the noun "door" in (1, 2 and 3) are considered a collocation, although they occur at various positions in the sentences.

(1) She *knocked* on his *door*.

(2) 100 women *knocked* on Donaldson's *door*.

(3) A man *knocked* on a metal front *door*.

Regardless of such differences between definitions, there is a consensus on the three main criteria for defining collocations: "non-compositionality", "non-substitutability" and "non-modifiability" (Manning and Schütze 2002:186)**.** According to the first criterion, the meaning of the collocation is not a straightforward composition of the meanings of its parts: either the meaning is completely different from the free combination – as in 'look out', which means 'to be careful' – or there is a connotation or added element of meaning that cannot be predicted from the parts (as in 'white wine' and 'white hair' where the 'white' color refers to slightly different colors) (Manning and Schütze 2002).

The second criterion in defining collocations (i.e. non-substitutability) stipulates that the components of a collocation cannot be substituted with other words even if they are synonymous. For instance, 'chance' cannot substitute 'probability' in 'high probability' although they are synonymous. Finally, non-modifiability means that many collocations cannot be modified with additional lexical items or through grammatical transformations. (Manning and Schütze 2002).

Collocations are among the widely used structural components in SNLP and they are the ones used for the present thesis as well. They are used for a variety of tasks such as Word Sense Disambiguation (WSD) (O'Hara et al. 2004), statistical Part-of-Speech (POS) Tagging (Lu 2005) and parsing (Mani 2004). Moreover, they are used for a variety of applications such as computational lexicography which tries to automatically identify the important collocations to be listed in a dictionary (Evert and Krenn 2003), natural language generation so as to make sure that the output sounds natural (Stone and Doran 1996) and Machine Translation (MT) since a word may be translated differently according to the collocation it occurs in (Smadja et al. 1996).

Collocations discovery is a non-trivial task because a collocation is not simply a frequent *n*-gram[10]. For example, function words are highly frequent words, yet they cannot be considered collocations as in "of the" (Manning and Schütze 2002). Moreover, the combination of two words can be frequent only because the two words are frequent like "next year" (Manning and Schütze 2002). Similarly, على أن/*ElY >n*/ (provided that) is not a collocation in Arabic but they are two function words occurring frequently together (Heliel 1989 as cited in Brashi 2005). Thus more accurate tools need to be used so as to make sure that the co-occurrence of two or more particular words is more than a chance (see section 2.4.2 below).

2.4.1.2 Parameters of Statistical Models

As previously defined, parameters are the variables of the structural components of a model. As for collocations, which are the structural components mentioned above, the variables can be considered in terms of the type of the parts of the collocation and/or their number.

[10] An *n*-gram is a sub-sequence of *n* items from a given sequence. An *n*-gram of size 1 is a *unigram*, size 2 is *bigram*, size 3 is *trigram* and size 4 or more is called an '*n-gram*' (McEnery & Wilson 2001).

According to Manning and Schütze (2002), there are different categories of collocations, such as compound nouns (e.g. *disk drive*), phrasal verbs (e.g. *make up*) and phrases (*bacon and eggs*, *weapons of mass destruction*). Other categories of collocations include "light verbs" collocations (Manning and Schütze 2002: 186), such as 'make a decision', 'do a favor' ... etc. Proper nouns are usually included in the category of collocations in NLP. Terminological expressions are also considered as collocations in spite of being fairly compositional. This is because they tend to be rather fixed and to be translated in fixed terms. Such categories are the possible parameters or variables for collocations as structural components.

2.4.1.3 Instantiation

Burch and Osborne (2003) state that instantiation is assigning probabilities to the parameters of the used structural component(s). It is an equivalent to *estimation*, which is discussed in detail in the following section.

In brief, modeling involves deciding on the structural component which is chosen according to the task under investigation. The parameters of such components are also to be decided on and afterwards the instantiation/estimation process starts.

2.4.2 Estimation

Estimation is the second core method in SNLP. Having the models constructed, and the parameters decided on, they should be estimated, that is, values should be assigned to such parameters. Collocations are usually estimated via "association measures".

Association measures are "mathematical formulas determining the strength of association between two or more words based on their occurrences and cooccurrences

in a text corpus" (Pecina and Schlesinger 2006: 651). There are many estimation methods some of which are briefly outlined in the following lines[11].

Conditional Probability (CP) is one of the easiest association measures, being a straightforward linear one. CP is the probability of an event given some knowledge of a previous event (Manning and Schütze 2002). It is formalized as:

$$\boldsymbol{P(x|y)} = \frac{P(x \cap y)}{P(y)}$$

Where
x is the event whose probability is to be calculated and
y is the given event

According to CP, the two events *x* and *y* are independent if $P(x \cap y) = P(x)$ (i.e. the probability of *x* is not affected by the existence of *y*).

Pointwise Mutual Information (PMI) is "the amount of information provided by the occurrence of the event represented by *y* about the occurrence of the event represented by *x*" (Manning and Schütze 2002: 179). It is formalized as:

$$\boldsymbol{PMI(x|y)} = \log_2 \frac{P(x|y)}{P(x)\,P(y)}$$

where
x is the event whose probability is to be calculated
y is the given event

According to Manning and Schütze (2002), PMI can be considered a good measure of independence, since values close to 0 indicate independence. However, it is a bad measure of dependence because it is biased to rare event.

T-test is another association measure which looks at the difference between the observed mean of a sample and the expected mean of the distribution scaled by the variance of the data[12] (Manning and Schütze 2002). It is formalized as:

[11] The association measures used here are the ones outlined in Manning and Schütze (2002)

[12] The mean is the average offset. The mean of a sample is denoted as *x* and the mean of the distribution as μ. The variance measures how much the individual offsets deviate from the means. It is represented as s^2 (Manning and Schütze 2002).

$$t = \frac{x - \mu}{\sqrt{\frac{s^2}{N}}}$$

where
x is the mean value of the sample
μ is the mean value of the distribution
s^2 is the sample deviation
N is the total size of the corpus

The interpretation of the T-test values is not straightforward. A distribution of T-test critical values must be used where a proportion value *p* is chosen and so are a percentage of performance *C* and a degree of freedom (d.f.). If the resulting value is lower than the d.f. chosen, then the two words are unlikely to collocate.

Pearson's chi-square (X^2) is defined as "the sum of the difference between the observed values and the expected values" (Manning and Schütze 2002: 169). It is computed as follows:

$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where
$\sum_{ij}$ is the summation of the probabilities of the observed events
O_{ij} are the probabilities of the observed events
E_{ij} are the probabilities of the expected events

Like T-test, the interpretation of X^2 results is not straightforward. A distribution table of X^2 critical values must be used so that a proportion value *p* is chosen and a corresponding d.f. is picked and used to compare the results, if the resulting value is less than the d.f. chosen, then the two words are unlikely to be a collocation and vice versa[13] (Manning and Schütze 2002).

Log-likelihood ratio is another association measure whose results are more interpretable than X^2: the higher the value is the more likely that the two words collocate. Log-likelihood ratio is calculated as follows:

[13] See Appendix A for the distribution tables of *T-test* critical values and X^2 critical values

$$\log\lambda = 2 \times [n11 \times \log(n11/m11) + n12 \times \log(n12/m12) + n21 \times \log(n21/m21) + n22 \times \log(n22/m22)]$$

Where n11 is the observed probability that the two words have occurred together
m11 is the expected probability that the two words have occurred together
n12 is the observed probability that the first word has without the second word
m12 is the expected probability that the first word has without the second word
n21 is the observed probability that the second word has without the first word
m21 is the expected probability that the second word has without the first word
n22 is the observed probability that neither the second word not the first has occurred
m22 is the expected probability that neither the second word not the first has occurred

Unlike both T-test and X^2, the results of log-likelihood ratio are more interpretable: high results indicate that the words are likely to collocate and vice versa. However, Moore (2004) argues that log likelihood ratio introduces a substantial amount of error when computing rare events.

Thus there are many association measures used for estimating the collocational relation between two or more existing words. Each measure has its pros and cons and each seems suitable for a particular study.

2.4.3 Corpora

SNLP is a corpus-oriented field (Jurafsky and Martin 2000, Burch and Osborne 2003), which requires huge amounts of corpora since it depends on actual examples observed in the corpus for its *learning* phase (see section 1.3.3). However, the availability of high-quality corpora appropriate for each SNLP task and application is a major problem in the field (see section 2.5).

2.4.4 Evaluation

Evaluation is the last method of SNLP, according to Burch and Osborne (2003). The evaluation method deals with the evaluation methodology and the evaluation metric(s), both of which are discussed below.

2.4.4.1 Evaluation Methodologies

Among the most widely used evaluation methodologies are the "training-and-testing paradigm" (Jurafsky and Martin 2000: 204) and the baseline models. In the

former, the corpus is divided into two sets: training and testing; the language model is built according to the training set and is evaluated against the testing set. There are many ways to manipulate the training and the testing sets according to the type and the amount of the data available. First, if there is enough data, the corpus can be divided as 80:20 or 90:10 for the training and the testing respectively (Jurafsky and Martin 2000).

Second, if there is a small amount of data available for training, it may be difficult to divide the data into training and testing sets because the testing set will fairly reflect the data as a whole. For such cases, cross-validation is usually used, where the data is randomly divided into *n* sections. The learner is trained from *n*-1 of these sections and then evaluated against the remaining one. This is done *n* times and the performance of the system is reported as the average of the *n* evaluations (Burch and Osborne 2003).

Finally, if there are different possible language models, and the best one is to be picked, the corpus is divided into three sections: "the training set", "the development test set" and "the test set" (Jurafsky and Martin 2000: 204). The development test set – also called a devtest set – is used to pick up the best model and to tune new parameters. When the best model is chosen, it is then evaluated against the true test set.

In the training-and-testing paradigm, the training corpus must be carefully designed so as to avoid bias, overfitting and underfitting (Burch and Osborne 2003). Overfitting is "the failure to generalize beyond the training data so as to handle new situations" (Burch and Osborne 2003: 8). Underfitting is over-generalization. Moreover, the testing corpus must be well-built so as to efficiently represent the data.

Baseline models are "the simplest implementation that one can think of" (Burch and Osborne 2003: 19). For instance, in POS tagging a baseline model can be a model that always assigns the most frequent tag given a particular word without attempting to do any contextual disambiguation. A baseline for AR could be choosing the closest noun that agrees in gender, number and rationality with the pronoun. Having a baseline model allows a reference point which can determine how good the performance is.

Whether the training-and-testing paradigm or a baseline model is used, the question is: how to identify the correct instances? Usually, a gold standard is used. A *gold standard* is a manually-crafted set of examples, against which the results are compared (Burch and Osborne 2003). Available annotated corpora can be good gold standard sets, yet in the absence of an adequate gold standard, the researcher might need to build one.

2.4.4.2 Evaluation Metrics

Evaluation metrics are measures to compute the success of a proposed algorithm. Among the most widely used evaluation metrics are precision, recall and the F-measure. **Precision** is defined as "a measure of the proportion of selected items that the system got right" (Manning and Schütze 2002: 268). It is calculated as follows:

$$\textbf{Precision} = \frac{\textit{number of correctly resolved events}}{\textit{number of correctly resolved events} + \textit{number of incorrectly resolved events}}$$

Recall is defined as "the proportion of the target items that the system selected" (Manning and Schütze 2002: 268). It is calculated as:

$$\textbf{Recall} = \frac{\textit{number of correctly resolved events} + \textit{number of incorrectly resolved events}}{\textit{total number of events}}$$

F-measure is the weighted mean of precision and recall. It is calculated as:

$$\textbf{F-measure} = \frac{2 \times (\textit{precision} \times \textit{recall})}{\textit{precision} + \textit{recall}}$$

Currently there are many studies to develop better evaluation methodologies and metrics. Regardless of the evaluation methodology or the metric used, Burch and Osborne (2003) emphasize that performing an error analysis, which identifies the errors of the algorithm and the reasons for them, leads to indications about how the system may be improved in subsequent redesigns. Performing an error analysis is one stage in the development of a SNLP system: being able to inspect the types of errors that are being made, and being able to generalize them into linguistic features is useful for the subsequent redesign of the statistical model used by the system.

2.5 Problems with SNLP

Since SNLP is corpus-oriented, its success largely depends on the quality and the availability of the appropriate data which is not necessarily available. Consequently, SNLP systems usually undergo the problem of the *sparseness of data* which means that many of the cooccurrences of the words are either not observed or observed only once even in very large corpora (Karov and Edelman 1996).

In other words, a corpus is *sparse* if it has a large number of zero-probability events that should really have some non-zero probability (Jurafsky and Martin 2000). Kilgarriff and Grefenstette (2003) among others consider corpora with millions of words as small data sets that contain only a sample of the dominant meanings and usage-patterns and where rare words and rare meanings of common words and

combinations of words have almost no evidence. In brief, sparseness of data is an inherent problem in SNLP systems.

There are two approaches to handle sparseness of data: a statistically-motivated approach and a linguistically-motivated one. The former relies on smoothing that is "the task of re-evaluating some of the zero-probability and low-probability *n*-grams and assigning them non-zero values" (Jurafsky and Martin 2000: 207). The main idea of smoothing is discounting the probability of seen events and giving even a low-probability to the unseen ones.

There are many smoothing techniques briefly discussed in the following lines[14]. The first one is Laplace's Add-One (Manning and Schütze 2002). According to this technique a space is given to unseen events using the following equation:

$$P_{lap}(w_1 \ldots w_n) = \frac{C(w_1 \ldots w_n) + 1}{N + B}$$

Where
C is the number of counts of the word *w*
N is the total number of vocabulary tokens
B is the total number of vocabulary types

According to Jurafsky and Martin (2000) and Manning and Schütze (2002), the problem with Laplace's Add-One is overestimation: it assigns too much probability to unseen events, that is, it is biased to unseen events.

Another smoothing technique is Lidstone's Law of Succession. Instead of adding one, Lidstone's Law adds a positive value λ, which is usually ½. The equation goes as follows:

$$P_{lid}(w_1 \ldots w_n) = \frac{C(w_1 \ldots w_n) + \lambda}{N + B\lambda}$$

Where *C* is the number of counts of the word *w*
N is the total number of vocabulary tokens
B is the total number of vocabulary types
λ equals ½

[14] The smoothing techniques outlines here are the ones discussed in Manning and Schütze 2002

Overestimation of unseen events is decreased with Lidstone's Law of Succession and is further decreased with Good-Turing. Instead of discounting using a stable value, Good-Turing counts the probability of unseen events as follows:

$$P_{GT}(w_1 \ldots w_n) = \frac{N1}{N}$$

Where
N1 is the number of counts of the word *w*
N is the total number of vocabulary tokens

However, many linguists (Kilgarriff and Grefenstette 2003 and Keller et al. 2003, among others) argue against using smoothing techniques and go for using more data (i.e. large corpora). In other words, they suggest that probabilistic models based on very large corpora are better than the models based on such statistical estimates as smoothing. Thus they support the linguistically-motivated approach, which depends on using massive corpora, to deal with the sparseness of data.

One way to get such massive corpora is using the Web as corpus (Kilgarriff and Grefenstette 2003). The Web is a free, instantly available source of immense amounts of documents, representing almost all possible languages and genres. Moreover, according to Kilgarriff and Grefenstette (2003), its documents are to be considered a corpus because McEnery and Wilson (2001) define the corpus as any collection of more than one text provided that it is sampling, representative, machine-readable and standard. Manning and Schütze (2002: 129) broaden the definition of the corpus, saying that it is simply "a certain amount of data from a certain domain of interest, without having any way in how it is constructed". According to these definitions, the Web and its documents are indeed a corpus.

Using the Web as corpus has many advantages. First, it helps avoid bias to a certain language genre or domain (Kilgarriff and Grefenstette 2003). Usually, the statistics of a language model change according to the type of texts used for building

it. This imposes a limitation on the applicability of any language model, because it can be applied to new texts that might not be of the same type of the texts involved in the language model. The only way to guarantee the good performance of a language model is to draw it from random samples from different language types and genres which might be available through using the Web as corpus.

Second, the Web is a good source for massive monolingual, bilingual and multilingual corpora. Not only can it be used to collect such types of corpora (Resnik 1999), but also bilingual Web search engines can be used to search for translations (Kilgarriff and Grefenstette 2003).

Third, Web counts are proved by Keller and Lapata (2003) to be reliable enough due to the high correlation between Web frequencies and corpus frequencies, the reliable correlation between Web frequencies and human plausibility judgments, and the reliable correlation between Web frequencies and frequencies recreated using class-based smoothing and the correlation with the counts derived from a well-balanced corpus.

Many studies are based on using the Web as corpus. For instance, Keller and Lapata (2003) examined how useful the Web is as a source of frequency information for rare items, especially for dependency relations. Moreover, Volk (2001) gathered lexical statistics for resolving Prepositional Phrase (PP) attachments. Villasenor-Pineda et al. (2003 as cited in Kilgarriff and Grefenstette 2003) balanced their corpus using Web documents and Mihalcea and Moldovan (1999 as cited in Kilgarriff and Grefenstette 2003) built a Word Sense Disambiguation (WSD) engine using hit counts to rank word sense frequencies. Other studies include Jones and Ghani (2000) who built a language-specific corpus using the Web from a single document in that

language and Fujii and Ishikawa (2000) who acquired counts of collocations from the Web.

However, the Web has its own limitations. One of the main problems of using the Web as corpus is that Web documents might be noisy (Kilgarriff and Grefenstette 2003) because they are produced by a wide variety of authors cheaply and rapidly with little concern for correctness. That is why spelling and grammatical mistakes are quite common in Web documents. For example, on Google search engine, a search for "I beleave" gets 3,910 search results, and "I beleive" gets 70,900; yet the correct "I believe" gets more than 4 million results. Consequently, Keller et al. (2002) suggest that the large amount of the data available for Web counts overcomes the problem of noisy texts.

Moreover, there are inherent problems in search engines themselves (Kilgarriff and Grefenstette 2003). First, the search engine results do not present enough instances as they are usually limited to 1000 results maximum. Second, they do not present enough contexts for each instance; Google provides ten-word fragments around the target word. Finally, the results always vary according to the search engine load (i.e. the number of documents uploaded to the search engine). However, in spite of such problems Web as corpus is still a promising approach (Kilgarriff and Grefenstette 2003).

In brief, sparseness of data is a problem with SNLP. It is statistically handled via smoothing techniques and linguistically handled through using the Web as corpus. To date, there are no clear results about which approach performs better, although some opinions support the linguistically-motivated approach.

2.6 Applications of SNLP

There are many applications for SNLP such as Question Answering (QA), Text Classification (TC), parsing, POS tagging and Machine Translation (MT). The following subsections briefly outline two applications – POS tagging and MT – that are directly related to the present thesis.

2.6.1 Part-of-Speech (POS) Tagging

Part-of-Speech (POS) tagging (Burch and Osborne 2003) is an early success of SNLP. The basic task is to assign a label – from a set of POS tags – to each token encountered. The most popular method to assign POS tags is to take a large corpus of sentences marked with tags and then train a model upon those tagged sentences. However, a range of other methods is also possible. Current taggers operate at around 96% per-token accuracy (Burch and Osborne 2003).

Burch and Osborne (2003) argue that POS taggers have such high accuracy for a number of reasons. First, closed-class words (i.e. function words) are usually unambiguous. Second, the per-word distribution of possible tags of the open-class words is usually sharply peaked around a few tags. This means that in most situations, there are only a few possible tags. Finally, the context required to disambiguate most words is usually only a few words and/or surrounding POS tags. For some non-English languages, current taggers are effective, but it is an open question whether they are equally effective for all languages.

2.6.2 Statistical Machine Translation (SMT)

MT has a statistical formulation (Brown et al. 1993). In SMT, if the translation is from Arabic into English, then every English string *e* is considered a possible translation of an Arabic string *a*. Every pair of strings {*a*; *e*} is assigned a probability P(*e*|*a*) (i.e. the probability that a translator, when presented with *a* will produce *e* as its

translation). According to such a probability, the SMT system is supposed to give the most suitable *e* string given a new *a* one.

Various resources can be used to estimate such a probability like bilingual sentence-aligned corpora and manually translated corpora. Recent work on SMT takes two directions: the first is to augment the translation model with more linguistically sophisticated information; the second is to cope with scarce linguistic resources with more sophisticated statistical techniques.

2.7 Summary

SNLP is an approach of NLP that focuses more on statistical techniques and unsupervised or semi-supervised approaches. Its main methods are modeling, estimation, corpora and evaluation. Being a data-driven field, SNLP undergoes the sparseness of data, which can be statistically and/or linguistically handled. There is an increasing interest in SNLP and its various applications such as POS tagging and SMT.

Part Two

Anaphora Resolution

2.1 Introduction

Part two contains two chapters both discuss more specifically the problem of Anaphora Resolution (AR). The fist chapter defines different types of anaphora in both English and Arabic and focuses on the definition of AR in particular. Besides, it provides a detailed comparison between Arabic and English pronominal systems. Such a comparison highlights the differences between the two pronominal systems that result in poor MT performance.

The second chapter outlines the different AR approaches including the discourse-based and corpus-based ones. Not only does the chapter refer to different AR approaches, but also it shows how these approaches are applied to some MT systems. Both chapters theoretically illustrate AR within the framework of NLP.

Chapter Three

Anaphora Resolution and Poor Machine Translation Output

3.1 Introduction

After a brief definition of anaphora and Anaphora Resolution (AR), the chapter justifies for the poor performance of some current MT systems in terms of Arabic AR. The differences between the English and Arabic pronominal systems, which are detailed in the present chapter, are the main reasons for such poor performance that is proved through some illustrated examples at the end of the chapter.

3.2 What is Anaphora? What is Anaphora Resolution (AR)?

Anaphora is defined as "the reference to an entity that has been previously introduced into the discourse" (Jurafsky & Martin 2000: 672). The entity, to which the expression refers, is typically called the referent. However, Mitkov (1999) and Bussmann (1996) make a distinction between the referent and the antecedent, which is the term commonly used in the computational literature of AR. The referent is the object or the state of affairs in the extralinguistic reality to which the referring expression refers, whereas the antecedent is the linguistic realization of this entity. Thus throughout the present thesis, the researcher uses the term antecedent.

Anaphora is divided according to two criteria: the type of the anaphor and the position of the antecedent. The first criterion divides anaphora into indefinite NP anaphora, definite NP anaphora, pronominal anaphora, *one* anaphora and demonstratives (Deoskar 2004, Jurafsky & Martin 2000, Mitkov 1999, Werth 1999). Pronominal anaphora (i.e. pronominalization) – which is the main focus of the present thesis – is realized by anaphoric pronouns as in (3.1), where the 3rd person feminine singular pronoun /*hA*/ (them) refers to الخدمات الصحية /*AlxdmAt AlSHyp*/ (health services).

(3.1) ... دعم الدولة للخدمات الصحية بدلا من تركها[15] ...

Transliteration:

/... DEm Aldwlp llxdmAt AlSHyp bdlA mn trkhA .../

Translation:

... the support of the state for health services instead of discarding them ...

The second criterion (i.e. the position of the antecedent) classifies anaphora into intrasentential and intersentential anaphora (Werth 1999). If the antecedent is outside the sentence boundary, where the anaphor exists, this is usually referred to as intersentential as in (3.2). However, if the antecedent occurs within the same sentence of the anaphor, this is known as intrasentential anaphora, as in (3.3).

(3.2) فاتن حمامة هي سيدة الشاشة العربية. إنها الاستاذة...

Transliteration:

/fAtn HmAmp hy sydp AlAp AlErbyp. AnhA AlAstAzp .../

Translation:

Faten Hamama is the "First Lady of the Arab cinema". She is the master …

(3.3) مأساة العولمة أنها...

Transliteration:

/m>sAp AlEwlmp AnhA …/

Translation:

The problem with globalization is that it is...

Anaphora Resolution (AR) is the process of determining the antecedent of a given anaphor (Deoskar 2004, Mitkov 1999, Mitkov 2001). AR is required in many

[15] All Arabic examples are extracted from *Al-Ahram Newspaper* corpus.

NLP applications such as Question Answering (QA), Information Extraction (IE), Text Summarization (TS), Machine Translation (MT) … etc. (Deoskar 2004, Mitkov 1999). Thus a successful end-to-end system requires a successful AR algorithm.

3.3 AR and Machine Translation (MT)

Good-quality MT systems presuppose understanding the source text which involves discourse processing. One of the important prerequisites for understanding the source text is the ability to disambiguate pronouns, i.e. pronominal AR (Mitkov 1996).

According to Mitkov (1996), some current MT systems can translate only single sentences. Once they have an input of two consecutive sentences, the result is often disappointing since – in most cases – the second sentence is translated as a completely isolated piece of information, without taking the preceding one into consideration.

In fact, practical experience show that even within the sentence level some current MT systems are unable to correctly resolve anaphora, especially when dealing with typologically different languages such as Arabic and English. Not only do such systems dealing with Arabic/English translation process sentences separately, but also they do not make up for the differences between the two languages that are the main reason for MT poor performance as discussed in section (3.5).

3.4 Differences between Arabic and English Pronominal Systems

Arabic and English are typologically different languages. Not only do they belong to different language families, since Arabic is a Semitic language whereas English is an Indo-European one, but also they have different syntactic, morphological and semantic properties that affect the output of MT systems dealing with them. The following subsections review the differences between the two

languages in terms of their pronominal systems and with special focus on 3rd person pronouns, which are the main focus of the present thesis.

3.4.1 The Number of 3rd Person Pronouns in Arabic and English

In English, there are only three 3rd person pronouns, namely *he, she* and *it* and their derivatives (Berk 1999, Jurafsky and Martin 2000). The Arabic pronominal system, however, contains a larger set of the 3rd person pronouns which are illustrated in table (3.1).

Arabic Pronoun	Transliteration	Translation
ه	/*h*/	(he, him, his)
ها	/*hA*/	(she, her, hers)
هما	/*hmA*/	(they, them, their, theirs)
هن	/*hn*/	(they, them, their, theirs)
هم	/*hm*/	(they, them, their, theirs)

Table (3.1): Arabic 3rd Person Pronouns Studied in the Present Thesis

Having a larger set of 3rd person pronouns is not the only difference between English and Arabic pronominal systems. There are also other differences in terms of grammatical case, morphology, gender and number which cause the poor performance of some current MT systems in terms of AR as the following subsections show.

3.4.2 Grammatical Case

In English, the grammatical case – determined by the grammatical function of the pronoun – affects the form of the pronoun (Berk 1999); a pronoun functioning as an object has a different form from the one functioning as a subject as in (3.4) below:

(3.4) *They* are talking about *their* new house which was given to *them* by the government.

In (3.4), the subject pronoun *they* has a different form from the object pronoun *them* and the possessive pronoun *their*. Table (3.2) shows the different grammatical cases

and forms of the 3rd person pronouns in English. The nominative case usually corresponds to the grammatical function of the subject, the accusative to the object and the genitive to the genitive function as the nouns after prepositions and possessives.

Grammatical Case	Grammatical Form
Nominative	He/She/It/They
Accusative	Him/Her/It/Them
Genitive	His/Her/Hers/Its/Their/Theirs

Table (3.2): Grammatical Cases and Forms of English 3rd Person Pronouns

As table (3.2) shows, all English 3rd person pronouns can be used in the three grammatical cases and in turn functions. However, this does not apply to Arabic 3rd person pronouns.

In the Arabic language, pronouns have three grammatical cases: nominative, accusative and genitive (Alhashemy 2000, Gaber 1980, Hasan 1999). The nominative case goes with such grammatical functions as the subject and المبتدأ /*Almbtd*>/[16] (the topic)[17]. The accusative corresponds to the object. The genitive – as in English – goes with nouns after prepositions and المضاف إليه /*AlmDAf* <*lyh*/ (the word annexed to another)[18]. However, each grammatical case – and thus function – is used with a limited subset of the 3rd person pronouns.

The pronouns هـ /*h*/ (him/his/it/its) and ها /*hA*/ (her/hers/it/its) are only used in the accusative and the genitive cases (Algilayyeny 2003, Gaber 1980, Hasan 1999). This includes examples like:

[16] /*Almbtd*>/ (the topic) is the first word in an Arabic nominal sentence.
[17] Translation of Arabic grammatical terms are adopted from Badawi et al. (2004)
[18] Annexation is the "collocation of two elements, usually both nouns, in a fixed and inseparable unit broadly conveying the sense of possession" (Badawi et al. 2004: 130)

(3.5) عن انتخابات1992 التي فاز فيها بيل كلينتون

Transliteration:

/*En AntxAbAt 1992 Alty fAz fyhA byl klyntwn*/

Translation:

… about the 1992 elections which Bill Clinton won …

(3.6) الهزيمة التي لقيها الفريق القومي

Transliteration:

/*Alhzymp Alty lqyhA Alfryq Alqwmy*/

Translation:

The defeat which the national team has encountered

In (3.5), ها /*hA*/ (them) is encliticized to the preposition في /*fy*/ (in) and thus it is used in the genitive case. In (3.6), ها /*hA*/ (it) is used as an object pronoun in the accusative case, being an enclitic to the verb لقي /*lqy*/ (encounter).

The rest of the pronouns – namely, هما /*hmA*/ (they/them/their), هم /*hm*/ (they/them/their) and هن /*hn*/ (they/them/their) – are used in all the three grammatical cases. In the following three sentences, the same pronoun هم /*hm*/ is used in the three grammatical cases of the nominative, accusative and genitive.

(3.7) هم الأطفال والنساء

Transliteration:

/*hm Al>TfAl wAlnsA'*/

Translation:

These are children and women

(A nominative case; topic)

(3.8) الخبراء الذين يكلفون بتنفيذ البرامج التي تسند اليهم

Transliteration:

*/AlxrA' Al*yn yklfwn btnfy* AkbrAmj Alty tsnd Alyhm/*

Translation:

The experts who are allocated to execute the programs for which they are responsible…

(A genitive case; encliticized to a preposition)

(3.9) وعدد من الفقراء يتقدمهم

Transliteration:

/wEdd mn AlfqrA' ytqdmhm/

Translation:

… and a number of the poor led by …

(An accusative case; object)

As such, Arabic 3rd person pronouns, unlike the English ones, are further subdivided into two groups according to their possible grammatical cases: pronouns used in the accusative and the genitive cases only and pronouns used in all the three grammatical cases.

This is not the only difference, though. In English, each grammatical case is marked with a different pronoun form as table (3.2) above shows. In Arabic, nevertheless, the same form can be used to represent different grammatical cases as in sentences (3.7), (3.8) and (3.9) above, where the same form of هم /*hm*/ is used in the three grammatical cases and occupies three different grammatical functions.

In spite of such a strict classification of the 3rd person pronouns in terms of their grammatical cases, the identification of the correct grammatical case is not easy. The difficulty rises from the fact that the word to which the pronoun is encliticized might be ambiguous as in حبه /*Hbh*/ where حب /*Hb*/ might be a noun meaning *love* or a

verb meaning *to love*. If حب /*Hb*/ (love) is a noun, ه /*h*/ (his) is used in the genitive case being encliticized to a noun; but if حب /*Hb*/ (love) is a verb, ه /*h*/ (him) is used in the accusative case.

3.4.3 Morphology

In English, pronouns are not usually cliticized to any other morpheme. The following sentences extracted from Berk (1999) illustrate this.

(3.10) *He* is a painter.

(3.11) *She* had known *him* for ages.

(3.12) The bodybuilder flexed *his* body.

In Arabic, pronouns are divided into three groups in terms of their manner of cliticization. The first group consists of two pronouns, namely هو /*hw*/ (he) and هي /*hy*/ (she), which are never used in the enclitic form. In Arabic, this group is known as الضمائر المنفصلة /*AlDmA}r AlmnfSlp*/ (non-enclitic pronouns) (Algilayyeny 2003, Gaber 1980, Hasan 1999).

The second group of Arabic pronouns is that of the enclitic pronouns. The group includes two pronouns: ـه /*h*/ (him, his) and ها /*hA*/ (her, hers), which are never used separately; they must be encliticized to a word – a noun, a verb or a preposition (Algilayyeny 2003, Gbr 1980, Hasan 1999) as in (3.13) and (3.14). These pronouns are known as الضمائر المتصلة /*AlDmA}r AlmtSlp*/ (enclitic pronouns).

(3.13) يطالب عنان اسرائيل بتغيير سياستها

Transliteration:

/yTAlb EnAn AsrA}l btgyyr syAsthA/

Translation:

Anan asks Israel to change its policy

(3.14) هذا الدور الذي لعبته الأمم المتحدة

Transliteration:

/*h*A Aldwr Alzy lEbth AlAmm AlmtHdp*/

Translation:

This role played by the United Nations

The last group contains the pronouns هما /*hmA*/ (they, them, their, theirs; dual, masculine and feminine), هم /*hm*/ (they, them, their, theirs; plural, masculine) and هن /*hn*/ (they, them, their, theirs; plural; feminine) which can be used either separately or as enclitics (Algilayyeny 2003, Gaber 1980, Hasan 1999).

The encliticization of the last group pronouns depends on their grammatical case/function. When these pronouns are used in the nominative case – especially in the grammatical function of المبتدأ /*Almbtd>*/ (topic), they are more likely to be used separately as in sentence (3.15) below. However, when they are in the genitive form, they are likely to be encliticized to the preceding noun/preposition as in (3.16). Similarly, when they are used in the accusative form, they are very likely to be encliticized to the preceding verb as in (3.17).

(3.15) إصابة المهاجمين و هم يحاولون الهرب

Transliteration:

/*<SAbp AlmhAjmyn w hm yHAwlwn Alhrb*/

Translation:

The attackers got injured during their attempt to escape

(3.16) الفلسطينيين سوف يحصلون علي حقوقهم

Transliteration:

/*AlflsTynyyn swf yHSlwn Ely Hqwqhm*/

Translation:

The Palestinians will get their rights

(3.17) إنه يريد أن ينزلهما من...

Transliteration:

/Anh yryd An ynzlhmA mn ..../

Translation:

He wants to get them off...

The main problem that results from the fact that some of the Arabic 3rd person pronouns are encliticized to other morphemes or even written without a space as in the case of هو /*hw*/ and هي /*hy*/ and the conjunctions و /*w*/ and ف /*f*/ is that of ambiguity, both syntactic and lexical. The letters of the enclitic pronoun can be either a part of the word or a pronoun. For instance, المهن /*Almhn*/ can be analyzed in two different ways: as the definite article /*Al*/ procliticized to the plural noun مهن /*mhn*/ (jobs) or as a noun الم /*Alm*/ (pain) encliticized to the 3rd person feminine plural pronoun هن /*hn*/ (their).

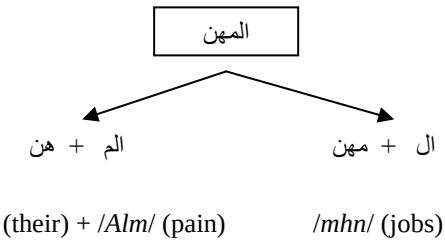

Diagram (3): Different Analyses of the Same Word Due to Pronoun Ambiguity

3.4.4 Gender

English 3rd person pronouns distinguish between male, female and "non-personal" genders (Berk 1999, Jurafsky & Martin 2000: 679) and between ±HUMAN entities. Therefore, the singular pronouns *she* and *he* are usually used for +HUMAN entities, whereas the singular pronoun *it* is usually used for –HUMAN entities. The following examples illustrate these points:

(3.18) I bought *a new computer* yesterday. *It* cost $1000.

(3.19) Sonia wants to marry *a tall man*; *he* must be handsome.

(3.20) *Sonia* wants to marry a banker; *she* insists.

In Arabic, words are either feminine or masculine. Feminine nouns are subdivided according to two criteria: animacy and the morphological form of the word. As for the first criterion, feminine words are مؤنث حقيقي /*m&nv Hqyqy*/ ("real feminine") or مؤنث مجازي /*m&nv mjAzy*/ ("unreal feminine") (Wright 1981:177). Real feminine words are necessarily animate entities referring either to humans like رانيا /*rAnyA*/ (Rania) or to animals like بقرة /*bqrp*/ (a cow). Unreal feminine words are always inanimate entities such as شجرة /*$jrp*/ (a tree) and ورقة /*wrqp*/ (a paper). Linguistically speaking, both real and unreal feminine words are the same; both stipulate feminine markers in verbs and are given the same feminine adjectives, demonstratives and referred to using the same feminine pronouns as in (3.21) and (3.22)

(3.21) هيلاري كلينتون وابنتها

Transliteration:

/*hylAry klyntwn wAbnthA*/

Translation:

Hilary Clinton and her daughter

(3.22) مأساة العولمة إنها...

Transliteration:

/*m>sAp AlEwlmp AnhA …*/

Translation:

The problem of globalization is that it is …

In (3.21) and (3.22), both هيلاري /*hylAry*/ (Hilary) and العولمة /*AlEwlmp*/ (the globalization) are feminine, even if the former is a real feminine noun and the latter is an unreal feminine one. Both are referred to using the same 3rd person feminine singular pronoun ها /*hA*/. However, English marks the two words differently, since *globalization* is –HUMAN, and therefore is referred to as *it*, and *Hilary* is +HUMAN and is thus referred to as *she.*

As for the second criterion, that is, the morphological form of the word, feminine words can be divided into three groups (Hasan 1999): المؤنث اللفظي /*Alm&nv AllfZy*/ ("feminine by form"), المؤنث المعنوي /*Alm&nv AlmEnwy*/ ("feminine by signification" and المؤنث اللفظي والمعنوي /*Alm&nv AllfZy wAlmEnwy*/ ("feminine by form and signification") (Wright 1981: 105, 246). The classification of these groups depends on the existence (or absence) of feminine markers. The feminine markers in Arabic are three: ة /*p*/ (*teh marbuta*) as in سفينة /*sfynp*/ (a ship), ى /*Y*/ (*alef maqsura*) as in ليلى /*lylY*/ (Laila: a feminine Arabic name) and اء /*A'*/ (*alef mamduda*) as in سماء /*smA'*/ (sky).

According to Hasan (1999), Alhashemy (2000) and Algilayyeny (2003), the first group of feminine words contains one of these feminine markers, in spite of being masculine in signification like حمزة /*Hmzp*/ (Hamzh: a masculine Arabic name) and أسامة /*>sAmp*/ (Osama: a masculine Arabic name). These words are dealt with as any other masculine word: they are given masculine verbs, masculine pronouns and masculine demonstratives as in (3.23).

(3.23) أدى حمزة واجباته

Transliteration:

/>dY Hmzp wAjbAth/

Translation:

Hamzh did his homework

The second group of feminine words – those that are feminine by signification – are feminine words that do not end in a feminine marker, yet they are feminine in meaning like شمس /*$ms*/ (sun). These are dealt with like any feminine word; they are used with feminine verbs, demonstratives and pronouns as in (3.24).

(3.24) شمس النهار بنورها الساطع

Transliteration:

/ *$ms AlnhAr bnwrhA AlsATE*/

Translation:

The morning's sun with its bright rays

The last group of feminine words, that is, feminine words by form and signification, includes words that end in one of the feminine markers as well as being feminine in signification like: منة /*mnp*/ (Menna: An Arabic feminine name). These nouns are just similar to all feminine words and they are referred to using feminine pronouns as in (3.25).

(3.25) استطاعت منة في فيلمها الجديد

Transliteration:

/*AstTAEt mnp fy fylmhA Aljdyd*/

Translation:

Menna could in her new movie

As for Arabic masculine words, they are only subdivided into two categories: real and unreal (Hasan 1999, Alhashemy 2000, Algilayyeny 2003). Real masculine words are animate entities referring either to HUMANs like أحمد />*Hmd*/ (Ahmed) or to animals like كلب /*klb*/ (a dog). Unreal masculine words are inanimate entities such

as كتاب /*ktAb*/ (a book). Linguistically speaking, both real and unreal masculine words are the same; both are given masculine verbs, adjective, demonstratives and referred to using masculine pronouns as in (3.26) and (3.27).

(2.36) استطاع أحمد نظيف خلال عمله...

Transliteration:

/*>stTAE >Hmd nZyf xlAl Emlh*/

Translation:

Ahmed Nazif managed through his work …

(3.27) فالكتاب يسلي صاحبه...

Transliteration:

/*fAlktAb ysly SAHbh*/

Translation:

The book entertains its owner

3.4.5 Number

According to Berk (1999), English nouns are singular, plural, non-count or collective nouns. Singular English nouns are usually modified by a number of determiners, most of which distinguish singular nouns from plural ones. Table (3.3) illustrates determiners that are specific to singular nouns which are referred to using the singular 3rd person pronouns of *he, she, it* in the nominative case; *him, her, it* in the accusative case; and *his, her(s), its* in the genitive case as in (3.28) and (3.29).

(3.28) I bought *a new computer* yesterday. *It* cost $1000.

(3.29) *Sonia* wants to marry a banker; *she* insists.

Determiner	An Example
A/An	A house
The	The photo
This	This child
That	That plate

Table (3.3): Determiners that Modify Singular English Nouns

English nouns are pluralized by adding plural suffixes as in table (3.4) or by internal vowel change as in *tooth, teeth* and *man, men*. Moreover, pluralization might involve vowel and consonant change as in *mouse, mice*. Some nouns in English have the same form in singular and plural like *deer* and *sheep*. Plural nouns are referred to using the plural 3rd person pronouns of *they*, *them*, *their* and *theirs* as in (3.30).

(3.30) I saw *three children* today. *They* were walking with *their* dog which was always following *them*.

The Plural Suffix	An Example
-s	Pens
-es	Boxes
-ies	Stories
-ves	Knives
-en	Children

Table (3.4): English Plural Suffixes

Non-count words in English are referred to using the same set of the singular 3rd person pronouns. However, unlike singular nouns, they are not used with the indefinite articles – a/an as in (3.31) and (3.32).

(3.31) *A cat* is sitting on the couch. *It* looks really nice.

(3.32) *Milk* is my favorite drink. I drink *it* every day.

Collective English nouns refer to collective entities with individual members like: *band, team, army* and *jury* (Berk 1999). If the speaker wishes to focus on the

group itself, the collective noun is usually referred to using a singular 3rd person pronoun as in (3.33). However, if the speaker wishes to focus on the individual members of the collectivity, the collective noun is likely to be referred to using a plural 3rd person pronoun like *they* and *their* in (3.34).

(3.33) *My team* is playing well this week. I hope *it* wins.

(3.34) Every debate *team* gets disqualified because *they* attack each other instead of attacking *their* opponents.

In Arabic, nouns can be singular, dual, plural, non-count or collective. Singular nouns are referred to using singular pronouns. The set of the Arabic 3rd person pronouns that refer to singular nouns includes هي /*hy*/, هو /*hw*/, هـ /*h*/ and ها /*hA*/.

Dual nouns – which do not exist in English – refer to two entities of the same type. Dual nouns are formed by adding one of two suffixes: either ان /*An*/ in the nominative case or ين /*yn*/ in the accusative and the genitive cases (Algilayyeny 2003). Dual nouns have their own 3rd person pronoun that refers to them in all possible grammatical cases, namely هما /*hmA*/ as in (3.35).

(3.35) الحرب بين الكوريتين و حلفائهما

Transliteration:

/*AlHrb byn Alkwrytyn w HlfA}hmA*/

Translation:

The war between the two Koreas and their alliances

As for plural nouns, they are divided into three groups: جمع المذكر السالم /*jmE Almzkr AlsAlm*/ (regular masculine plural), جمع المؤنث السالم /*jmE Alm&nv AlsAlm*/ (regular feminine plural) and جمع التكسير /*jmE Altksyr*/ (broken plural) (Hassan 1999, Alhashemy 2000, Algilayyeny 2003). Regular masculine plural works only on real

masculine nouns provided that they are +HUMAN and that they are not feminine by form. It is formed by adding two suffixes: either ون /*wn*/ in the nominative case or ين /*yn*/ in the accusative and the genitive cases (Algilayyeny 2003) as in (3.36).

(3.36) الحوار مفتوح للمثقفين بمختلف مشاربهم

Transliteration:

/AlHwAr mftwH llmvqfyn bmxtlf m$Arbhm/

Translation:

The conversation is open for all the cultured with their different interests

Regular feminine plural is formed by adding the suffix ات /*At*/ in all grammatical cases. As shown in table (3.5), different types of feminine words can be pluralized using the regular feminine plural form. The pronoun used to refer to regular feminine plurals is هن /*hn*/, given that the words refer to +HUMAN entities; whereas ها /*hA*/ is the pronoun used to refer to –HUMAN regular plurals (Hassan 1999).

(3.37) وإن السيدات اللاتي صادفتهن

Transliteration:

/w<n AlsydAt AllAty SAdfthn/

Translation:

And the women whom she met

The Singular Word	Its Type	Its plural form
أم />*m*/ (a mother)	Real	أمهات />*mhAt*/
مشكلة /*m$klp*/ (a problem)	Unreal	مشكلات /*m$klAt*/
حمزة /*Hmzp*/ (Hamzh)	Feminine by form	حمزات /*HmzAt*/
مريم /*mrym*/ (Mary)	Feminine by signification	مريمات /*mrymAt*/
شجرة /*$jrp*/ (a tree)	Feminine by form and signification	شجرات /*$jrAt*/

Table (3.5): Examples of Plural Feminine Arabic Words[19]

Broken plural is the third form of pluralization in Arabic. It is formed in three ways: adding an infix (e.g. سهم /*shm*/ (an arrow), سهام /*shAm*/ (arrows)), omitting a letter (e.g. رسول /*rswl*/ (a messenger), رسل /*rsl*/ (messengers)) or changing the internal structure of vowels through changing diacritics (e.g. أسَد />*asad*/ (a lion), أُسُد />*usud*/ (lions)) (Algilayyeny 2003: 191).

The pronoun that refers to a broken plural noun is determined by the singular form of the plural word (Hasan 1999). If the singular word is a +HUMAN masculine noun, the plural form is referred to by هم /*hm*/ as in (3.38). If the singular word is a +HUMAN feminine noun, the plural form is referred to by هن /*hn*/ as in (3.39). If the singular word is a –HUMAN noun, the plural form is referred to by ها /*hA*/ as in (3.40).

(3.38) قبول أوراق الطلاب بحد أدني 75% و ستعلن نتائجهم...

Transliteration:

/*qbwl AwrAq ALTlAb bHd Adny 75% w stEln nTA}ghm* /

Translation:

… Admitting the students' papers with a minimum score of 75% and their results will be announced...

[19] The examples are the researcher's

(3.39) النسوة في الخليج العربي يغسلن شعورهن بـ...

Transliteration:

/Alnsp fy Alxlyj AlErby ygsln $Ewrhn b…/

Translation:

Women in the Arabian Gulf wash their hair with...

(3.40) المعاهد التي تدرس الطب وفي مقدمتها

Transliteration:

/AlmEAhd Alty trds AlTb wfy mqdmthA/

Translation:

The institutions teaching medicine and first of them...

Non-count and collective nouns also exist in Arabic. As for the former, they are handled in the same way a singular noun is handled as in (3.41) where the non-count Arabic noun اللبن */Allbn/* (the milk) is referred to using the 3rd person singular masculine pronoun ه */h/*. Similarly, collective nouns are referred to by singular pronouns as in (3.42) where the 3rd person singular pronoun ه */h/* is used to refer to the collective Arabic noun فريق */fryq/* (team).

(3.41) معلومات هامة عن اللبن ومكوناته ومنتجاته

Transliteration:

/mElwmAt hAmp En Allbn wmkwnA wmntjAth/

Translation:

Important information about milk, its ingredients and its products

(3.42) احتفل فريق الأهلي بفوزه على الفريق المغربي**...**

Transliteration:

/AHtfl fryq Al>hly bfwzh ElY Alfryq Almgrby/

Translation:

Al-Ahly team celebrated its victory over the Moroccan team...

3.4.6 A Summary of Arabic Pronouns

It might be useful here to summarize the properties of the Arabic pronominal system. Table (3.6) summarizes such a system in terms of grammatical case, morphology, the gender and number of the pronoun, the number and the gender of its referents and the possible English translations of the pronoun.

Pronoun	Grammatical Case	Morphology	Gender	Number	Referent's Gender	Referent's Number	Possible English Translations
هي /*hy*/	Nominative	Separate	Feminine	Singular	Feminine: real, unreal, by signification, by form and signification	Singular	She
هو /*hw*/	Nominative	Separate	Masculine	Singular	Masculine –real, unreal- And feminine by form	Singular	He
هـ /*h*/	Accusative, Genitive	Enclitic	Masculine	Singular	Masculine – real, unreal – And feminine by form	Singular	Him/his It/its
ها /*hA*/	Accusative, Genitive	Enclitic	Feminine	Singular	Feminine: real, unreal, by signification, by form and signification	Singular, broken plural	Her/hers It/its
هما /*hmA*/	Nominative, Accusative, Genitive	Separate, Enclitic	Masculine, Feminine	Dual	Feminine: real, unreal, by form, by signification, by form and signification. Masculine: real, unreal	Dual	They/them/their
هم /*hm*/	Nominative, Accusative, Genitive	Separate or Enclitic	Masculine	Plural	Masculine Real	Regular Plural	They/them/their
هن /*hn*/	Nominative, Accusative, Genitive	Separate or Enclitic	Feminine	Plural	Feminine Real	Regular Plural	They/them/their

Table (3.6): A Summary of Arabic 3rd Person Pronouns

3.5 Examples of Poor Output of Some Current MT Systems

Due to the aforementioned differences between Arabic and English pronominal systems, the performance of some current MT systems is rather poor in terms of AR. The thesis does not attempt any kind of evaluation of any of the MT systems referred to, but the output of such systems mainly motivates writing the present thesis. Examples on such poor performance are given in the following lines:

(3.43) وصفت الخارجية الأمريكية الانتخابات الرئاسية فى مصر بأنها

خطوة إيجابية للأمام

Transliteration:

/wSft AlxArjyp Al>mrykyp AlAntxAbAt Alr}Asyp fy mSr b>nhA xTwp <yjAbyp ll>mAm/

Correct Translation:

The US Foreign Ministry described the Egyptian presidential elections as a positive step...

Sakhr's Translation:

The American Foreign Ministry described the presidential elections in Egypt that *it* is a positive step forward.

Systran's Translation:

Describer the external American presidential elections in Egypt that *she* positive step

In (3.43), the pronoun ها */hA/* in بأنها */b>nhA/* refers to الانتخابات الرئاسية */AlAntxAbAt Alr}Asyp/* (the presidential elections). Thus it should have been translated as *they* not as *it*.

(3.44) صرحت السيدة قرينة الرئيس بأنها ستكثف عملها بالتعاون ....

Transliteration:

/SrHt Alsydp qrynp Alr}ys b>nhA stkvf EmlhA bAltEAwn /

Correct Translation:

The First Lady announced that she will focus on cooperating with...

Sakhr's Translation:

The Mrs. announced the president's wife that *it* will intensify *its* work in cooperation …

Google's Translation:

Mrs. Suzanne Mubarak, the President stated that *it* will intensify *its* cooperation

In (3.44), ها */hA/* is used twice: first it functions as the subject of *that-clause*; second it is used in the genitive case as a possessive pronoun. In both cases, Sakhr (a dictionary-based MT system) and Google (a SMT system) are unable to correctly resolve it, although there is only one possible antecedent – السيدة قرينة الرئيس /*Alsydp qrynp Alr}ys*/ (The First Lady) – which is clear in terms of number and gender because of the suffix ة /*p*/ which is a marker of singular feminine words in Arabic (Alhashemy 2000, Hassan 1999).

(3.45) أكد رئيس مجلس الوزراء أنه نتج عن كل ذلك استقرار سعر العملة

Transliteration:

*/>kd r}ys mjls AlwzrA' >nh ntj En kl *lk AstqrAr sEr AlEmlp/*

Correct Translation:

The Prime Minister confirmed that all this resulted in stabilizing currency exchange

Sakhr Translation:

The Prime Minister confirmed that *he* resulted from all that a stability in the currencies price.

Systran Translation:

Head council of the ministers confirmed that he all that currency stability price resulted about

In sentence (3.45), ه */h/* is a non-pleonastic pronoun; it is used in an impersonal expression where the pronoun does not have a referent. However, both Sakhr and Systran (A SMT system) are unable to detect such a non-anaphoric usage of the pronoun, thus they yield wrong translations. These examples reflect the relatively poor output of some current MT systems, which is one of the main motivations of the present thesis.

3.6 Summary

Anaphora Resolution (AR) is an important step for efficient NLP applications such as Machine Translation (MT). MT systems dealing with the typologically different languages of Arabic and English usually handle AR insufficiently. This is attributed to the differences between the two languages in terms of gender, number, grammatical cases and morphology.

Chapter Four

Approaches to Anaphora Resolution

4.1 Introduction to AR Approaches

Approaches to AR can be divided according to the kind of knowledge used as input into discourse-based approaches, syntax-based approaches, hybrid approaches and statistical, corpus-based approaches. Discourse-based approaches, unlike others, are usually knowledge-rich ones that require too much human intervention. They depend on such theories as Centering Theory (CT) and Discourse Representation Theory (DRT) that require comprehensive details for the structural properties of the sentences as well as pragmatic knowledge about the real world which is usually hard to encode (Lappin 2005).

Syntax-based approaches vary from knowledge-rich approaches such as Hobbs (1977 as cited in Mitkov 1999) and Lappin and Leass (1994) to knowledge-poor ones like Williams et al. (1996) and Kennedy and Boguraev (1996). The knowledge-rich syntax-based approaches rely on fully parsed input, which might require human intervention. Knowledge-poor approaches, however, depend on partially parsed input, augmented with little semantic knowledge.

Hybrid and corpus-based approaches – which are the focus of the present thesis and which are further elaborated in section (4.3) – are unsupervised or semi-supervised knowledge-poor approaches that usually rely on statistical techniques supported with the least semantic and/or syntactic knowledge available.

Regardless of the type of knowledge used, all AR approaches rely on *constraints and preferences* (Carbonell & Brown 1988, Deoskar 2004, Jurafsky & Martin 2000, Lappin 2005, Mitkov 1999). These are divided into two classes: eliminating and preferential (Mitkov 1997). Eliminating constraints are "eliminating [features for] discounting certain noun phrases from the set of possible candidates",

whereas preferential constraints are "preferential [features that] give more preference to certain candidates and less to others" (Mitkov 1999: 3).

Eliminating constraints include a number of "hard-and-fast" features that exclude certain Noun Phrases (NPs) from being possible antecedents (Jurafsky & Martin 2000: 678). The most commonly used eliminating constraint is number and gender agreement. Number agreement is "a categorization of pronouns with respect to number"; only antecedents that agree in number with the pronoun are accepted as possible antecedents (Jurafsky & Martin 2000: 678). Gender agreement stipulates that the antecedent agrees with the gender specified by the pronoun. That is why اسرائيل */AsrA}yl/* (Israel) is the appropriate antecedent for ها */hA/* in (4.1).

(4.1) يطالب عنان اسرائيل بتغيير سياستها

Transliteration:

/yTAlb EnAn AsrA}yl btgyyr syAsthA/

Translation:

Anan asks Israel to change its policy

Although it is important for a practical NLP system to include a reasonably comprehensive set of such semantic constraints, such lexical knowledge is usually "difficult to encode in a comprehensive manner" (Jurafsky & Martin 2000: 681), besides being time and effort consuming (Lappin 2005).

Preferential constraints often come to play when an anaphor is ambiguous even after applying the eliminating constraints. "Recency", "grammatical role", "repeated mention", "grammatical parallelism" and "verb semantics" are all preferential constraints frequently used in the literature of AR (Jurafsky & Martin 2000: 681-694). The following lines specifically focus on recency which is used for the present thesis.

Most approaches to AR incorporate recency which is the notion that entities introduced in recent utterances are more salient than those introduced in utterances further back. Therefore, in (4.2) the pronoun ه /*h*/ is more likely to refer to الموقف /*Almwqf*/ (the attitude) than to شيئاً /*$y}F*/ (something).

(4.2) نرى عند الطهطاوي شيئاً جديداً يضيفه نظرياً إلي الموقف الذي وقفه

Transliteration:

*/nrY End AlThTAwy $y}A jdydA yDyfh nZryA Aly Almwqf Al*y wqfh /*

Translation:

We see that Al-Tahtawy has something new that he theoretically adds to his attitude

In brief, the type of the AR approach is determined on the basis of the knowledge incorporated in the algorithm. Discourse-based, syntax-based, corpus-based and hybrid are all AR approaches that share one main concept which is constraints and preferences. The following subsections elaborate on hybrid and corpus-based approaches discussing their main concept of knowledge-poor AR.

4.2 Knowledge-Poor AR Approaches

Knowledge-poor AR approaches rely on the input features which can be identified without reference to deep semantic/syntactic information or detailed real-world knowledge (Lappin 2005). The attempt to avoid the labor-intensive task of developing knowledge bases and the need to develop fully language-independent and potentially robust systems are the main motivations for such approaches (Mitkov 1998, Deoskar 2004). Moreover, relying on knowledge-poor approaches might speed up a given system, because not too much preprocessing is required (Lappin 2005).

However, such approaches undergo the problem of resources efficiency and coverage (Lappin 2005). Moreover, one opinion against knowledge-poor approaches is that they are genre-specific, since they were applied to restricted texts only (Lappin 2005). Nevertheless, there are many attempts to apply them to unrestricted texts like Sandra et al. (1996) who try a shallow syntactically-based technique to AR within the context of Information Extraction (IE) and Text Summarization (TS). Their study mainly focuses on English and achieves 61% accuracy. Similarly, Baldwin (1997) applies the same approach to narrative texts achieving 90% accuracy.

Other researchers like Deoskar (2004) think that it is "unfair" to compare between knowledge-rich and knowledge-poor approaches (Deoskar 2004: 10). This is because the former usually use manually preprocessed input data, whereas the latter are usually end-to-end systems that automate all the preprocessing stages; inaccuracies in the preprocessing stage lead to an overall reduction in the performance of a system. Examples on knowledge-poor approaches are given in the following subsections:

4.2.1 William et al. (1996)

One of the early attempts of knowledge-poor approaches is that of Williams et al. (1996). They develop a syntactic, rule-based system to handle AR in the context of TS and IE. Their system works on unrestricted English texts which belong to a vast number of genres, achieving 76% accuracy. They use a POS statistical tagger, an NP shallow parser, a set of rules and knowledge-bases of names, titles and general knowledge. Their system resolves pronominal and definite NP anaphora types.

4.2.2 Kennedy and Boguraev (1996)

Kennedy and Boguraev's (1996) system does not require in-depth or full syntactic parsing, because it works on the output of a POS tagger, enriched only with

annotations of the grammatical functions of the lexical items in the input text. After the morphological and syntactic filters are applied, the remaining set of candidate antecedents is subjected to the preferential constraints of recency and grammatical parallelism. The candidate with highest salience weight is determined to be the actual antecedent; in the event of a tie, the closest candidate is chosen. The approach works for both lexical anaphors (reflexives and reciprocals) and pronouns. Evaluation involves a random selection of genres, including press releases, product announcements, news stories, magazine articles, and other documents existing on World Wide Web (WWW) pages. Their system reports 75% accuracy.

4.2.3 Baldwin's COGNIAC (1997)

CogNIAC is a system developed at the University of Pennsylvania to resolve pronouns with limited knowledge and linguistic resources (Baldwin 1997). The system requires sentence detection, POS tagging, simple NP recognition and basic semantic category information for the preprocessing phase. CogNIAC is built on the following core rules (Baldwin 1997: 39-40):

1. Unique in discourse: if there is a single possible antecedent in the read-in portion of the entire discourse, then pick it as the antecedent
2. Reflexive: pick the nearest possible antecedent in the read-in portion of current sentence if the anaphora is a reflexive pronoun
3. Unique in current and prior sentence(s): if there is a single possible antecedent in the prior sentence and the read-in portion of the current sentence, then pick it as the antecedent
4. Possessive pronoun: if the anaphor is a possessive pronoun and there is a single exact string match of the possessive in the prior sentence, then pick it as the antecedent

5. Unique current sentence: if there is a single possible antecedent the read-in portion of the current sentence, then pick it as the antecedent
6. If the subject of the prior sentence contains a single possible antecedent, and the anaphor is the subject of the current sentence, then pick it as the antecedent
7. Pick the most recent: pick the most recent potential antecedent in the text

In COGNIAC, pronouns are resolved from left to right in the text. For each pronoun, the rules are applied in the presented order. For a given rule, if an antecedent is found, then the appropriate annotations are made to the text and no more rules are tried for that pronoun, otherwise the next rule is tried. If no rules resolve the pronoun, then it is left unresolved. The system reports 92% for precision and 64% for recall (Baldwin 1997).

4.2.4 Mitkov et al. (1998)

The knowledge-poor approach of Mitkov et al. (1998) is one of the most important approaches, being applied to more than one language – English, Arabic and Polish – and achieving the highest performance rate. However, it is only applied to technical manuals, which are syntactically and lexically restricted and thus they are less challenging than unrestricted texts.

The approach takes as an input the output of a POS tagger, identifies the NPs which precede the anaphor within a distance of 2 sentences backwards, checks candidates for gender and number agreement and then applies the so-called antecedent indicators to the remaining candidates by assigning a positive or negative score. The NP with the highest aggregate score is proposed as antecedent.

The core of the approach lies in activating the empirically-based antecedent indicators which play a decisive role in tracking down the antecedent from a set of possible candidates. These indicators are definiteness/indefiniteness, givenness, indicating verbs, lexical reiteration, section heading preference, non-prepositional NPs, relative pronouns, collocations, immediate reference, sequential instructions, referential distance and preference of terms.

The approach is evaluated against a corpus of technical manuals (223 pronouns) and achieved a success rate of 89.7% for English, 95.2% for Arabic and 93.3% for Polish.

4.3 Corpus-Based AR Approaches

Corpus-based AR approaches are knowledge-poor approaches that focus more on probabilities and statistical techniques. They are either unsupervised or semi-supervised approaches as exemplified in the following subsections.

4.3.1 Dagan and Itai (1990)

Dagan and Itai (1990) performed an experiment to resolve references of the pronoun *it* in sentences randomly selected from the corpus. The model uses co-occurrence patterns observed in the corpus as preferential constraints. Candidates for antecedents are substituted for the anaphor and only those candidates available in frequent co-occurrence patterns are approved. They report an accuracy of 87%.

4.3.2 Ge at al. (1998)

Ge et al. (1998) use a small training corpus from the Penn Wall Street Journal Treebank marked with coreference resolution. They obtain an accuracy of 65.3% using just recency and syntactic constraints. After adding word information to the model – gender, number and animacy – the performance rises to 75.7%. Adding information about "mention count" – i.e. the more frequent a referent is, the more

likely it is to be the correct antecedent – improves accuracy to the final value of 84.2%.

4.3.3. Soon et al. (2001)

Soon et al. (2001) develop a system not only for AR but also for the resolution of all definite descriptions. They use a small annotated corpus to obtain training data to create feature vectors. These training examples are then given to a machine learning algorithm to build a classifier and a decision-tree-based algorithm. Their system is an end-to-end system which includes sentence segmentation, POS tagging, morphological processing, NP identification and semantic class determination.

Their feature vector consists of twelve features to indicate the type of the NP, to capture the distance between an anaphoric NP and its coreferent and to handle such features as gender, number, location, time, date money ... etc. The algorithm achieves a recall rate of 58.6%, precision of 67.3% and an F-measure of 62.6%.

4.3.4 Ng and Cardie (2002)

Ng and Cardie (2002) try to make up for the lack of linguistically-based features in Soon et al. (2001) and add syntactic and lexical features. As a result, their system achieves a performance rate of 70.4%.

4.3.5 Uryupina (2006)

Uryupina (2006) investigates the usability of linguistically-based features for statistical AR. The linguistic features of similarity, semantic compatibility and salience are integrated into a statistical model for AR. According to her results, such features reduce error rate by 19.9%.

4.4 AR Approaches and MT Systems

After reviewing relevant AR approaches to the present thesis, the following lines review their application to MT systems. All approaches are applied to a variety

of languages within the framework of MT, yet none of them is applied to the Arabic/English MT systems to the best of the researcher's knowledge.

As cited in Mitkov (1999), discourse-based AR approaches that rely on the Discourse Representation Theory (DRT) and the Binding Theory were applied to an English/Japanese MT system (Wada 1990) and an English/Chinese MT system (Chen 1992). Statistical, corpus-based approaches are also applied to Japanese/English MT systems to resolve Japanese zero pronouns, using such semantic preferences as conjunctions and verbal semantic attributes to determine intrasentential antecedents of Japanese zero anaphors with a success rate of 98% (Nakaiwa et al. 1995). Similarly, Preuß et al. (1994) work on a statistically-based AR approach for the English/German MT system KIT-FAST, making use of proximity, binding, parallelism and conceptual consistency. In spite of these studies, none is applied to Arabic/English MT systems.

4.5 Conclusion

This chapter briefly outlines AR approaches. It focuses on knowledge-poor and statistical, corpus-based approaches which are directly related to the proposed AR algorithm. Both approaches rely on the least available morphological, syntactic and/or lexical knowledge. However, corpus-based approaches give more attention to probabilities and statistical techniques than the knowledge-poor ones.

Part Three

Corpus Preprocessing

3.1 Introduction

This part handles the corpus preprocessing phase in three chapters. The first two chapters provide the theoretical background of the corpus preprocessing tasks used in Arabic NLP in general and Arabic AR in particular. The first chapter deals with such general corpus preprocessing tasks as punctuation markers, numbers, Out-Of-Vocabulary (OOV) words, diacritics, normalization, code-switching and the required level of word analysis. The second chapter, however, deals with more Arabic AR-specific corpus preprocessing tasks such as light stemming, semantic features acquisition and non-pleonastic pronouns identification. Both chapters set a theoretical framework for the corpus preprocessing tasks used in the present thesis.

The last chapter of this part elaborates on the researcher's methodology in handling all required corpus preprocessing tasks. In addition to using off-the-shelf tools such as tokenizers and POS taggers, the researcher develops an Arabic AR-specific tokenization scheme and a set of special tools to handle semantic features acquisition and non- pleonastic pronouns identification. Each of these points is detailed in the following chapters.

Chapter Five

The Tokenization Scheme

5.1 Introduction

In Computational Linguistics (CL), the tokenization scheme determines what type of units is to be preprocessed and then submitted to the algorithm as input. In this sense, tokenization is defined as "the identification of orthographically valid string units that can be submitted … for analysis" (Buckwalter 2002: 3). In other words, tokenization is the initial "preprocessing step" concerned with "identifying the basic units to be processed" (Webster and Ku 1992: 1106).

Identifying such units relies mainly on the task considered; they might be single words, phrases, complete sentences, or even idioms and fixed expressions. That is why Habash (2005) emphasizes that there is no one single possible or obvious tokenization scheme. Moreover, he defines a tokenization scheme as "an analytical tool devised by the researcher" so as to serve his/her research purposes (Habash 2005: 578).

Many issues are considered in a tokenization scheme. Such issues include punctuation markers, numbers, unknown words, diacritics, normalization, code-switching and the required level of word analysis. The subsequent sections illustrate how each point is dealt with in Arabic CL.

5.2 Punctuation Markers

According to Jurafsky and Martin (2000), dealing with punctuation markers as words depends on the task and on how much information such markers give to the task under study. There are tasks such as grammar-checking, spelling-error detection, author identification and POS tagging where punctuation is counted as words, yet there are other tasks that discards the existence of punctuation markers such as Word Sense Disambiguation (WSD).

According to Badawi et al. (2004), written Modern Standard Arabic (MSA) adopts and adapts Western punctuation. However, in MSA, unlike English, the positioning of punctuation is determined more by rhetorical and acoustic factors than by the content of the phrases and sentences. Consequently, Badawi et al. (2004) state that punctuation markers in MSA texts are usually inconsistent.

Chalabi (2001: 518) emphasizes that "[a] major problem faced in handling Arabic computationally is the rare use of punctuation markers" in most of current available corpora. The same idea is emphasized by Ali (1993: 25), as he states that "punctuation in English follows strict rules [but] Arabic punctuation is much more flexible, and its usage is rather discretional".

Studies in Arabic Natural Language Processing (ANLP) have two approaches to handle punctuation markers. First, in the case that the used corpus has a rather systematic punctuation system; punctuation markers are separated and used as words, phrases or sentences delimiters (Larkey and Connell 2002, Sadat and Habash 2006 among others). Second, in the case that the used corpus inconsistently uses punctuation markers, they are completely removed from the corpus (Sarkar and Roeck 2004 among others).

To sum up, keeping or removing punctuation markers depends on the used corpus, on how much information these markers provide and on the task under study.

5.3 Numbers

Like punctuation markers, keeping or removing numbers depends on the studied task and the information they provide. Like punctuation markers, ANLP studies either remove or keep numbers as delimiters. Khoja (2001) and Sarkar and Roeck (2004) remove them, whereas Sadat and Habash (2006) use them to mark word boundaries. One of the few studies that develops a different approach to deal with

numbers is that of Nelken and Shieber (2005). They normalize numbers to one single tag, namely "NUM".

5.4 Out-Of-Vocabulary (OOV) Words

Out-Of-Vocabulary (OOV) words – also known as "unanalyzed words" – are "the words that receive no analysis from the morphological analyzer" (Habash 2005: 576). OOV words are a problem frequently encountered in ANLP.

Khoja (2001) encounters the same problem while building her Arabic Part-Of-Speech Tagger (APT). Up to 21% of the words in her testing corpora – ≈ 85,000 tokens extracted from different Arabic newspapers – are OOV words. Out of such OOV words, 67% are foreign proper nouns. Thus Khoja (2001) decides to tag all OOV words as nouns. Unlike Khoja (2001), OOV words in Habash (2005) are removed being only 0.5% of the testing corpora. However, Habash (2005) agree with Khoja (2001) that OOV words are frequently proper nouns. Unlike Khoja (2001) and Habash (2005), Diab et al. (2004) treat OOV words like any other tag giving them their own probabilities on the basis of their existence in the training corpora. In Diab et al.'s (2004) POS tagger, OOV words are tagged as NOFUNC (i.e. No Function).

5.5 Diacritics

Diacritics are the short disambiguating vowels used in written Arabic (Badawi et al. 2004). They include the following short vowels: (ـَـ) الفتحة /*AlftHp*/ (*Fateha*), (ـِـ) الكسرة /*Alksrp*/ (*Kaserah*), (ـُـ) الضمة /*AlDmp*/ (*Dammeh*), (ـْـ) السكون /*Alskwn*/ (*Skun*) and (ـّـ) الشدة /*Al$dp*/ (*Shaddah*). However, one of the main characteristics of written MSA is the absence of diacritics (Badawi 1973).

The absence of diacritics has two effects on MSA used in the written media language. First, it results in augmenting ambiguity, both syntactic and lexical. For example, the non-diacriticized كتب /*ktb*/ can be read as كَتَبَ /*katab*/ (he wrote), as كُتُب

/*kutub*/ (books), كَتّبَ /*kat~ab*/ (he dictated), or as كُتِبَ /*kutib*/ (it was written) (Chalabi 2001, Beesley 2001, Sadat and Habash 2006 among others).

Moreover, Badawi (1973) states that the absence of diacritics in MSA has resulted in "fixing word order since syntactic-semantic relationships between words cannot be deduced without diacritics" (Badawi 1973: 145). In Arabic, the subject is always in the nominative case, which is marked in the singular word forms with the diacritic (ـُـ) الضمة /*AlDmp*/ (*Dammeh*), and the object is always in the accusative case, marked in singular word forms with (ـَـ) الفتحة /*AlftHp*/ (*Fateha*). As a result, in Arabic the object can precede the subject, being diacritically marked as in (5.1).

(5.1) ضرب غلامَه زيدُ[20]

Transliteration:

/*Drb glAmahu zydu*/

Translation:

Hit his boy Zeid = Zeid hit his boy

However, due to the absence of diacritics in MSA, word order tends to be rather fixed (Badawi 1973).

Due to the absence of diacritics in most of written MSA texts, many studies in ANLP decide to remove them such as Abdelali (2005), Sadat and Habash (2006) and Nwesri et al. (2007) among others.

5.6 The Level of Word Analysis

Generally speaking, there are two levels of word analysis in ANLP: the root and the stem. According to Darwish and Oard (2002), the root is the linguistic unit of

[20] This is an archaic example adopted from Hassan (1999). The reason of using such an archaic example is the fact that such structures do not exist in written MSA due to the absence of diacritics.

meaning, which has no prefixes, suffixes or infixes; whereas a stem is the morphological base of a word to which affixes can be attached to form derivatives.

There is almost a consensus in ANLP applications that the root is "semantically too ambiguous to be practically useful" (Habash 2004: 1). For instance, the root of أولاد /*>wlAd*/ (sons/ children) is the tri-consonantal word ولد /*wld*/, out of which many words with different lexical meanings and syntactic categories can be derived as illustrated in diagram (4). As a result, many ANLP studies such as (Larkey et al. 2002 among others) use the lightly stemmed Arabic word form, given that it is less ambiguous than the root.

وَلَد /*walad*/ (a boy, sing. n.)
وُلِدَت /*wulid*/ *wulidat*/ (was born, pass. v.)
وَلَدَت /*waladat*/ (she gave birth to, past v.)
وَالِد /*waAlid*/ (a father/ a parent, sing. n.)
وَالِدَة /*waAlidap*/ (a mother, sing. n.)
وَلِيد /*walyid*/ (newly born, adj./ a baby, sing. n.)
تَوَالُد /*tawaAlud*/ (generating, sing. n.)
مَوْلِد /*mawolid*/ (birthday, sing. n.)
مِيلَاد /*miylaAd*/ (birthday, sing. n.)

Diagram (4): Different Derivations from the Arabic root ولد/wld/

Light stemming, when applied to Arabic, refers to the "process of stripping off a small set of prefixes and/or suffixes, without trying to deal with infixes, or recognize patterns and find roots" (Larkey 2002: 276). The process involves stripping off "a rich system of affixation and clitics" (Habash and Rambow 2005: 573), which is summarized in table (5.1). An Arabic word may have $0 \leq 2$ proclitics and $0 \leq 1$ enclitics.

Arabic Clitics			
Proclitics		**Enclitics**	
The Proclitic	**The Proclitic Type**	**The Enclitic**	**The Enclitic Type**
ال /Al/ (the)	Definite Article	ي /y/ (my)	1st singular person pronoun
بـ /b/ (by/with)	Preposition	نا /nA/ (we/us/our)	1st plural person pronoun
كـ /k/ (as)	Preposition	كما /kmA/ (you/your)	2nd dual person pronoun
لـ /l/ (for/to)	Preposition	كم /km/ (you/your)	2nd plural masculine person pronoun
و /w/ (and)	Conjunction	كن /kn/ (you/your)	2nd plural feminine person pronoun
فـ /f/ (so/then)	Conjunction	ه /h/ (him/his)	3rd singular masculine person pronoun
		ها/hA/ (her/hers)	3rd singular masculine person pronoun
		هما/hmA/ (they/them/their/theirs)	3rd dual person pronoun
		هم /hm/ (they/them/their/theirs)	3rd plural masculine person pronoun
		هن /hn/ (they/them/their/theirs)	3rd plural feminine person pronoun

Table (5.1): Arabic Affixation System

5.7 Normalization

Spelling Normalization – also known as orthographic normalization (Sadat and Habash 2006) – conflates orthographic variations. Usually, spelling normalization is the solution for the inconsistency of the corpus itself. Three letters pose problems in Arabic orthography, *alef*, *ya'* and *teh marbuta* (Larkey et al. 2002 and Xu et al. 2002 among other names).

As for the letter *alef*, Buckwalter (2004) shows that there are many orthographic variations of *alef* such as the writing (or the omission) of *hamza* above or below *alef* in stem-initial position, the writing (or the omission) of *madda* on *alef*

also in stem-initial position and the relatively free interchange of stem-initial *hamza above alef* and *hamza below alef*. Moreover, the use of *bare alef* without *madda* or *hamza* instead of all such variations is quiet common.

Although Arabic poses rather strict rules on writing *alef* and *hamza*, MSA newspapers and newswire do not seem to follow such rules (Buckwalter 2004). This is revealed in the variation of *alef* in *Al-Ahram Newspaper*. For instance, the word أسبانيا is frequently written with a *bare alef* as اسبانيا /*AsbAnyA/* (Spain), and equally it is written as أسبانيا />*sbAnyA/* (Spain) with a *hamza above alef*. However, a less frequent spelling is إسبانيا /<*sbAnyA/* (Spain) with a *hamza below alef*.

Different ANLP tools have different methods to deal with the normalization of *alef*. For instance, Buckwalter's (2002) AraMorph treats instances of *hamza above alef* and *hamza below alef* as they occur in the corpus; that is, if the word is written with a *hamza above alef* it is given only the analyses relevant to such an *alef* variation and no alternatives of the *alef* are introduced. The same thing happens when the word is written with *hamza below alef*; only relevant analyses are given. However, when the word is written with either a *bare alef* or *madda on alef*, the analyzer gives the word all its possible analyses, taking into consideration all possible orthographic variations of *alef*. Thus the word is analyzed in three different ways: as if written with a *hamza above alef*, and then analyzed again as if written with a *hamza below alef*, and finally as if written with a *bare alef*.

In some systems (e.g. Diab et al. 2004), the normalization of *alef* is not a problem since the POS tagger follows a statistical approach. Other ANLP systems chose to normalize all variations of *alef* to the *bare alef* (Sadat and Habash 2006 among others).

As for *ya'*, there are two variations: the dotted *ya'* ي /*y*/ and the non-dotted *ya'* ى /*Y*/ known as *alef maqsura*. Seemingly, Arabic newswire does not consider such a difference, although the two letters are completely different in the Arabic alphabet. Thus it is likely to find typologically incorrect words such as متي /*mty*/ (when) and حتي /*Hty*/ (until) in *Al-Ahram Newspaper* although they are correctly written as متى /*mtY*/ and حتى /*HtY*/. As a result of such variations, ANLP studies typically normalize *ya'* just like *alef* (Khoja 2001, Sadat and Habash 2006 among others). Buckwalter (2002), however, chooses to give his system a Second-Lookup.

In Buckwalter's (2002) Second Lookup, the word is looked up according to the form in which it is written; whether this form ends in *ya'* or *alef maqsura*. If the first look up does not yield any results, the *ya'* is converted into *alef maqsura* or vice versa, and the word is looked up again.

5.8 Concatenation

According to Buckwalter (2004), concatenations are run-on words. The most frequent in Arabic newswire is the combinations of function words *la* and *ma* with perfect or imperfect verbs such as لايزال /*lAyzAl*/ (still), مازال /*mAzAl*/ (still) and with nouns such as in لاشك /*lA$k*/ (no doubt) and لابد /*lAbud*/ (definitely). Moreover, proper nouns, especially those involving the word عبد */Ebd/* like عبدالله /*EbdAllh*/ and عبدالناصر /*EbdAlnAsr*/ are written either separately or in concatenated forms.

Such words do not pose a problem for ANLP tools. Buckwalter's (2002) AraMorph handles them however they are written: if they are written separately, it deals with each word separately giving it a separate POS tag and morphological analysis; if they are written in the concatenated form, then they are given a single POS tag and one morphological analysis as a proper noun. Moreover, Diab et al.'s (2004) tools are statistically developed and thus they handle the corpus however it is written.

5.9 Code-Switching

According to Bussmann (1996), code-switching is the switch of language varieties within communicative situations, whether monolingual, bilingual or multilingual. Code-switching is used to mark relationships, signal status and situation, deference and intimacy. Using slang and Classical Arabic words in *Al-Ahram Newspaper*, whose main language variety is MSA, is considered an instance of code-switching. Sometimes slang words or classical ones are used to mark informality or to enforce formality, respectively.

There is a kind of intersection between MSA and slang on one hand, and between MSA and Classical Arabic on the other hand. Slang is characterized by "the innovative use of common vocabulary" (Bussmann 1996: 437). Badawi (1973) states that MSA is relatively affected by Classical Arabic. Therefore, it is rather expected to find common vocabulary between the three language varieties of MSA, slang and Classical Arabic. Such intersection makes slang and Classical Arabic vocabulary difficult to be distinguished from MSA vocabulary. For instance, the word ساعتها /*sAEthA*/ in (5.2) is slang meaning 'at that time' but in (5.3) it is an MSA word meaning 'her watch'.

(5.2) إذا انتصرنا في الحرب ساعتها سنحب هذا الوطن أكثر

Transliteration:

/*<*A AntSrnA fy AlHrb sAEthA snHb h*A AlwTn >kvr*/

Translation:

If we won the war then, we would love this home more

(5.3) استطاعت أخيراً أن تسترد ساعتها

Transliteration:

/*AstTAEt >xyrA >n tstrd sAEthA*/

Translation:

She could finally get her watch back

5.10 Summary

In the previous sections the researcher tried to present a brief review of the main points that are tackled in a tokenization scheme. Such a review sets the theoretical background for some decisions discussed in chapter 7 which deals with the tokenization scheme used for the present thesis.

Chapter Six

Corpus Preprocessing Tasks

6.1 Introduction

A real-world AR algorithm vitally relies on the efficiency of such preprocessing tools as: POS taggers, morphological analyzers, NP chunkers, parsers ... etc. in order to analyze its input. Inaccurate preprocessing leads to a "drop" in the performance of the algorithm however accurate it might be (Mitkov 2001: 111). Inaccuracy is not the only problem of such preprocessing tasks, but for some languages such as Arabic they might not be even available.

As for "AR task-specific preprocessing tools", such as non-pleonastic pronouns identifiers and animacy and gender taggers, they usually receive less attention than such aforementioned "standard preprocessing tools" which are constantly developed and improved (Mitkov 2001: 115). That is why the performance of such AR task-specific preprocessing tools is still "far from ideal" (Mitkov 2001: 114).

As far as the Arabic language is concerned, preprocessing problems are more complicated. This is because such AR task-specific preprocessing tools are unavailable. Moreover, the performance of such standard preprocessing tools as POS taggers, morphological analyzers and parsers is not efficient (Khoja 2001, Buckwalter 2002, Onaizan and Knight. 2002, Diab et al. 2004 and Habash and Rambow 2005 among others).

As a result of the above limitations, in terms of preprocessing tools, the majority of AR algorithms do not operate in a fully automatic mode (Mitkov 2001). human intervention can be on a large-scale as in the manual stimulation of the approach itself or on a smaller-scale as in the cases where the evaluation samples are manually stripped of non-pleonastic pronouns (Dagan and Itai 1990, Kennedy and Boguraev 1996); where the output of the POS tagger and the NP chunker is manually

post-edited (Mitkov 1998) or where the output of the parser is manually corrected as in Lappin and Leass (1994).

To sum up, AR preprocessing handles a number of issues including: tokenization, POS tagging, semantic features acquisition, non-pleonastic pronouns identification, named-entity recognition and NP chunking. The success of an AR algorithm largely depends on the accuracy of the preprocessing tools and their availability in the first place. For some languages like Arabic, which is known for its scarce NLP tools and resources (Diab et al. 2004), many of such preprocessing tools either do not exist or their performance still needs to be largely improved (Habash 2007). Each of the following subsections provides a review of literature for each preprocessing tool in terms of the target language, namely Arabic.

6.2 Tokenization (Light Stemming)

In general, stemming equates or conflates variant forms of the same word into equivalent classes. When applied to Arabic, *light stemming* refers to a process of stripping of a set of prefixes and suffixes (i.e. clitics)[21] without trying to deal with infixes or recognize patterns and roots (Larkey et al. 2002). There are many Arabic stemmers/tokenizers used for different NLP tasks. What follows is a brief review of some of the freely available Arabic stemmers/tokenizers.

6.2.1 Some Current Arabic Stemmers/Tokenizers

6.2.1.1 Khoja (2001)

While building her Arabic Part-of-Speech Tagger (APT), Khoja (2001) develops her own stemmer which achieves an accuracy rate of 97% using a dictionary of 4,748 trilateral and quadrilateral roots. According to Khoja's (2001), her stemmer has two main problems which lower its performance. First, some of the letters that

[21] See table (5.1) for a list of Arabic clitics.

appear to be affixes are in fact parts of the words. Another problem is with the adjustment rules; that is, some letters may change to other letters when an affix is added, thus the letters should be changed when the affix is removed.

6.2.1.2 Diab et al. (2004)

Diab et al. (2004) develop a Support Vector Machine (SVM) tokenizer, which is one of the most widely used Arabic tokenizers being freely available and highly accurate. SVMs are supervised learning algorithms that rely on annotated training data, taken – in Diab et al. (2004) – from the Arabic TreeBank[22]. Tested on Arabic TreeBanks themselves, the tokenizer achieves an F-measured performance rate of 99.12%

6.3 POS Tagging

POS tagging is the process of assigning a label (from a set of POS tags) to each token encountered (Burch and Osborne 2003). There are different approaches for POS tagging and many Arabic POS taggers. However, there are some difficulties.

One main problem for Arabic POS taggers is ambiguity – lexical, morphological and syntactic ambiguity. Such ambiguity results from the absence of disambiguating diacritics in written Modern Standard Arabic (MSA) (Badawi et al. 2004, Freeman 2001, Habash 2004) and the Arabic rich system of affixation and clitics (Habash 2004). When POS taggers are used as an intermediate phase for another task, which is the case in AR algorithms, ambiguity will be a real problem. The following subsections illustrate some current Arabic POS taggers and how they deal with ambiguity.

[22] Arabic Penn TreeBank 1 (v2.0) is a MSA corpus containing Agence France Presse (AFP) newswire articles ranging from July through November of 2000. The corpus comprises 734 news articles covering various topics such as sports, politics, news, etc (Diab et al. 2004).

6.3.1 Some Current Arabic POS Taggers

6.3.1.1 Khoja's (2001) APT

Khoja (2001) develops a hybrid Arabic POS Tagger (APT), using statistical and rule-based techniques. She depends – whether in her rules or in her tagset – on the traditional Arabic grammatical theory not on Indo-European based rules and tagsets. The tagger is trained on 50,000 MSA words extracted from *Al-Jazirah Saudi Newspaper* in order to build a lexicon of 9,986 lightly-stemmed word types.

As for POS disambiguation, APT follows a statistical approach making use of two probabilities: the "lexical probability" which is "the probability of a word having a certain tag" and the "contextual probability" which is "the probability of one tag following another" (Khoja 2001: 5). For instance, contextual probabilities, trained on the previously mentioned training corpus, show that the probability of a verb being followed by a noun is 92.6% which is higher than the probability of having a noun followed by another noun (7.11%). Measured on ambiguous words, the disambiguation procedure achieves an accuracy rate of 90% (Khoja 2001).

APT yields 90% accuracy because the lexicon does not cover all possible tags of some words (Khoja 2001) due to the small training set. For instance, حرس /*hrs*/ (guards/ he guarded) appears in the training corpus as a noun meaning (guards) but not as a verb meaning (he guarded). Khoja (2001) suggests adding the missed tags manually or using a larger training corpus. However, Khoja's (2001) APT has one important advantage that is not realized in other Arabic POS taggers: it tags semantic features.

6.3.1.2 Buckwalter's (2002) AraMorph

AraMorph is a sophisticated rule-based POS tagger and morphological analyzer developed by Buckwalter (2002). It uses a concatenative lexicon-driven

approach where morphotactics and orthographic rules are built directly into the lexicon itself. In spite of being a widely used POS tagger and morphological analyzer, there are no reported results about its performance rates.

The system consists of three components: the lexicon, the compatibility tables and the analysis engine. The lexicon consists of three parts: a prefixes lexicon, a stems lexicon and a suffixes lexicon. The compatibility tables specify which morphological categories are allowed to co-occur. The analysis algorithm is rather simple since all of the hard decisions are coded in the lexicon and the compatibility table.

AraMorph tags the morphologically-marked semantic features. For instance, the suffix ات /*At*/ is encliticized to plural, feminine nouns and thus the Arabic noun كتابات /*ktAbAt*/ (writings) is tagged by AraMorph as NOUN, FEM and PL. The same thing goes for other indicative suffixes like ة /*p*/ which usually indicates feminine, singular nouns and ون /*wn*/ that indicates masculine, plural nouns (Hassan 1999, Alhashemy 2000). However, if the word is not attached to any indicative suffixes like شمس /*$ms*/ (sun: a feminine word in Arabic), it is not tagged for any semantic features.

One main problem with AraMorph is that of ambiguity. The system does not contain a POS disambiguation module and thus it renders all possible analyses of a given word. Having 135 distinct morphological labels, the system might give up to ten POS tags for a single word. This makes it rather problematic when used as an intermediate phase for more complicated NLP tasks. For example, while using it to create Arabic Penn TreeBank 1 (v2.0), the correct POS tag is manually chosen (Diab et al. 2004); and this is an expensive process. The same problem might encounter any AR algorithm that uses AraMorph.

6.3.1.3 Diab et al. (2004)

Diab et al. (2004) develop a Support Vector Machine (SVM) approach to automatically tokenize, POS tag and annotate base phrases in Arabic texts. SVM is a supervised learning approach that has the advantage of being robust where it can handle a large number of (overlapping) features. Arabic Penn TreeBank 1 (v2.0) is the corpus used by Diab et al. (2004) as a training corpus. According to standard evaluation metrics, the SVM-POS tagger achieves an accuracy rate of 95.49%, given that it handles the problem of ambiguity statistically based on the probabilities of the training corpus.

Diab et al. (2004) show that 50% of the errors encountered result from confusing nouns with adjectives or vice versa. This is to be expected because of the inconsistencies of the training data. For example, the word *United* in *United States of America* or *United Nations* is randomly tagged as a noun, or an adjective in the training data.

According to Diab et al. (2004), the results of the SVM-POS tagger are comparable to the state-of-the-art results of English texts when trained on similar-sized data. However, it does not tag semantic features since it annotates the segmented words, resulting from the tokenization module, using the Arabic Penn TreeBank POS tagset that does not include semantic features. The absence of such features affects some higher NLP applications such as AR.

Thus in the previous lines, some Arabic POS taggers are reviewed. There are two reasons for choosing these taggers among others: first, they are among the most recent taggers; second, they are almost the only available ones for public use. Khoja's (2001) has a low coverage rate, yet it tags semantic features. Buckwalter's (2002) is a morphological analyzer whose main task is to provide all possible analyses of a given

word. Finally, Diab et al.'s (2004) is the highest in performance, yet it does not tag semantic features that are basic features for certain NLP tasks such as AR.

6.4 Semantic Features Acquisition (SFA)

The semantic features of gender, number and rationality are important for two reasons. First, they are among the main differences between English and Arabic pronominal systems[23]. Due to such differences, some current MT systems poorly handle AR as in (6.1) where two of the current MT systems – Sakhr and Google – mistranslate the pronoun ها /*hA*/ (she/her).

(6.1) صرحت السيدة قرينة الرئيس بأنها ستكثف عملها بالتعاون ...

Transliteration:

/*SrHt Alsydp qrynp Alr}ys b>nhA stkvf EmlhA bAltEAwn*/

Sakhr's Translation:

The Mrs. announced the president's wife that *it* will intensify *its* work in cooperation with …

Google's Translation:

Mrs. Suzanne Mubarak, the President stated that it will intensify its collaboration...

Second, such features are one of the most widely used semantic constraints in the literature of AR[24] (Jurafsky and Martin 2000, Mitkov 1998, Williams et al. 1996, Kennedy and Boguraev 1996 among others). Even purely syntactic approaches (Lappin and Leass 1994), statistical approaches (Dagan and Itai 1990) and knowledge-poor ones (Mitkov 1998) make use of them. This thesis is no exception:

[23] See section 3.4 about the differences between Arabic and English pronominal systems and their effect on the performance of some current MT systems.
[24] See section 4.1 about using number and gender agreement as filters for candidate antecedents in AR algorithms.

the semantic features of gender, number and rationality are also used as semantic constraints.

However, such basic semantic features are not fully covered by current Arabic POS taggers. As mentioned in section (6.3), current Arabic POS taggers and morphological analyzers either avoid them completely like Diab et al. (2004) or tag them only in the case that they are morphologically marked like Buckwalter (2002). Even the taggers that tag such semantic features like Khoja's (2001) do not have good performance rates. Consequently, the researcher develops an approach to bootstrap them monolingually and bilingually (see section 7.4.3).

6.5 Non-Pleonastic Pronouns Identification

According to Badawi et al. (2004), Arabic recognizes non-pleonastic pronouns which are non-anaphoric pronouns that are usually invisible in translation. One example of non-pleonastic pronouns, according to Badawi et al. (2004) is (6.2) where the pronoun ه /*h*/ (he/him/his) encliticized to the particle أن /*>n*/ (Indeed) disappears in the English translation:

(6.2) أشار إلى أنه تم تكليف فريق عمل ...

Transliteration:

/<$Ar >lY <nh tm tklyf fryq Eml/

Translation:

He pointed to the fact that the commissioning of a working group had been completed.

In Arabic grammar theory, a non-pleonastic pronoun can be a congruent pronoun, a binding pronoun, an anticipatory pronoun, a separating pronoun or a pronoun encliticized to a verb following a relative pronoun. The "congruent pronoun" ضمير المشابهه /*Dmyr Alm$Abhh*/ (Badawi et al. 2004: 312) is used to separate الخبر

/*Alxbr*/ (the predicate) of الجملة الإسمية /*Aljmlp Al<smyp*/ (the nominal sentence) from a possible demonstrative as in (6.3).

(6.3) هذا هو سر التقدم الغربي.

Transliteration:

*/h*A hw sr Altqdm Algrby/*

Translation:

This is the secret of the western progress

The "separating pronoun" ضمير الفصل */Dmyr AlfSl/* (Badawi et al. 2004: 342) as in (7):

(6.4) الموقف الذي يتخذه الرئيسان ... هو موقف إيجابي

Transliteration:

*/Almwqf Al*y ytx*h Alr}ysAn hw mwqf >yjAby/*

Translation:

The stance which the two presidents are taking … is a positive one.

The "anticipatory pronoun" ضمير الشأن */Dmyr Al$>n/* (Badawi et al. 2004: 337) is encliticized to a nominal sentence modifier followed by a verbal sentence (Badawi et al. 2004: 320). Nominal sentence modifiers are a set of particles known as إن وأخواتها */>n w>xwAthA/* ("Inna and its sisters"). The set includes إن/ أن (Indeed: for emphasis), لكن (But: for contrast), كأن (as if: likeness), لعل (perhaps: probability, expectation and hope) and ليت (wish: for wishing). One example of anticipatory pronouns is:

(6.5) أكد أنه في المرحلة المقبلة سيكون التركيز على مراكز الطب الوقائي

Transliteration:

/>kd >nh fy AlmrHlp Almqblp sykwn Altrkyz ElY mrAkz AlTb AlwqA}y/

Translation:

He stressed the fact that in the next stage the concentration would be on preventative medical centers

The "binding pronoun" ضمير الربط */Dmyr AlrbT/* (Badawi et al. 2004: 327) is used to bind الخبر الجملة */Alxbr Aljmlp/* (the compound predicate) of the nominal sentence to its topic المبتدأ */Almbtd>/*. The compound predicate can be a nominal, a verbal or a prepositional phrase as in the following:

(6.6) هل تعرف أن العالم به 700 مليون سيارة؟

Transliteration:

/hl tErf >n AlEAlm bh 700 mlywn syArp/

Translation:

Do you know that there are 700 million cars in the world?

(6.7) هذه المهنة تواجهها عدة صعوبات.

Transliteration:

*/h*h Almhnp twAjhhA Edp SEwbAt/*

Translation:

A number of difficulties face this profession

(6.8) الطبيبة الأمريكية لايبدو على وجهها أي تعبير

Transliteration:

/AlTbybp Al>mrykyp lAybdw ElY wjhhA >y tEbyr/

Translation:

No expression shows on the American doctor's face

Finally, a non-pleonastic pronoun can be the pronoun encliticized to a verb after a relative pronoun as in (6.9) (Badawi et al. 2004: 491):

(6.9) الصعوبات التي واجهها ...

Transliteration:

/AlSEwbAt Alty wAjhhA/

Translation:

The difficulties he faced

As for the purposes of this study, congruent and separating pronouns are irrelevant; the thesis focuses on enclitic pronouns, whereas these pronouns are usually separate. The only kind of regularity that can be computationally implemented is the pronouns encliticized to verbs after a relative pronoun.

To the best of the researcher's knowledge, no previous studies address Arabic non-pleonastic pronouns identification. Even studies that deal with Arabic AR (Mitkov 1998) discard such a point.

6.6 Summary

Anaphora Resolution (AR) requires many preprocessing tools such as tokenization, POS tagging, Semantic Features Acquisition (SFA) and non-pleonastic pronouns identification. Previous parts give a brief review of literature for such tools in terms of Arabic NLP in general.

Chapter Seven

Corpus Preprocessing Methodology

7.1 Introduction

This chapter deals with the researcher's corpus preprocessing methodology in terms of the tokenization scheme and the AR-specific corpus preprocessing tools. After defining the used corpus, the chapter discusses the tokenization scheme and gives details about the decisions being made for each point. AR-specific corpus preprocessing tasks are elaborated afterwards in terms of methodology and performance rates.

7.2 Corpus Description and Counting

The used corpus is *Al-Ahram Newspaper* from 1998 to 2006 that consists of 42,538 files of various topics such as sports, politics, economics …etc. The language variety used in *Al-Ahram Newspaper* is the Modern Standard Arabic (MSA), which is the variety of Arabic used in most written media, news reporting, and some television talk shows in the Arabic-speaking world. It has its special characteristics that distinguish it from Classical Arabic. Such characteristics include: preference to NPs, despite the fact that VPs are also commonly used, and the absence of diacritics (Badawi 1973).

As for corpus counting, the *wordform* token and type counts are used. The token count is "the total number of running words in the corpus … [where] each word is counted each time it occurs", whereas the type count is "the total number of different words in the corpus [where] each word is counted only once irrespective of how often it occurs" (Olohan 2004: 200). According to Jurafsky and Martin (2000), wordform token and type counts deal with words as they appear in the corpus with all possible affixes and clitics.

Using wordform counting minimizes human effort, keeping the corpus as natural as possible. However, it leads to counting في /*fy*/ (in), وفي /*wfy*/ (and in) and فيهم

/*fyhm*/ (in them) as separate tokens and types, although they are different wordforms of the same stem في /*fy*/ (in). Nevertheless, this is irrelevant to Anaphora Resolution (AR). According to Olohan (2004), types and tokens identification and ratio are important in such studies aiming at author style identification, comparing the vocabulary size of two corpora of the same size or at translation studies. Since the scope of the thesis falls outside such studies, it is not really important if the different wordforms of the same stem are counted as different tokens and types. Therefore, wordform token and type counts are used for corpus counting. Accordingly, *Al-Ahram Newspaper* corpus has ≈ *2,000,000* wordform tokens and ≈ *971,000* wordform types.

7.3 The Tokenization Scheme

As mentioned earlier in chapter 5, a tokenization scheme is "an analytical tool devised by the researcher" so as to serve his/her research purposes (Habash and Rambow 2005: 578). The main objective of any tokenization scheme is to separate "the input stream into a graph of words" (Grefenstette et al. 2005). The following subsections describe the researcher's tokenization scheme used to define 'the word' in the present thesis. This tokenization scheme deals with:

1. Punctuation Markers.
2. Numbers.
3. Out-Of-Vocabulary (OOV) words.
4. Diacritics.
5. Level of Word Analysis.
6. Normalization.
7. Concatenation.
8. Code-switching.

Generally, keeping, removing and/or changing any units in the corpus depend on how much information they give to the AR algorithm. Therefore, the present tokenization scheme is not *the* scheme but *a* scheme that is devised to serve the purposes of the present thesis.

There are two objectives that the current tokenization scheme achieves in the present thesis. These are:

1. Minimizing the number of possible candidate antecedents. The more specific the used units are, the fewer candidates the algorithm has and therefore the easier its task is. However, an important point to be taken into consideration is that minimizing the number of possible candidates does not necessarily result in minimizing ambiguity.
2. Minimizing human intervention so that the corpus is left as natural as possible; the less human effort, the more economical.

The following subsections discuss each point in the current tokenization scheme, and how they are tackled to achieve the aforementioned objectives.

7.3.1 Punctuation Markers

As mentioned in (5.2), keeping or removing punctuation markers depends on the type of the used corpus and how much information they provide. Run-on lines, in *Al-Ahram Newspaper*, are frequently used: a complete paragraph might not contain any punctuation markers except for a final full-stop to mark the paragraph boundary as in paragraph (7.1) below:

(7.1) وأكد المسعودي رفض نقابة الصحفيين للحكم باعتباره سابقة جديدة في قضايا النشر وطالب بوضع حد نهائي له بعد أن شغل الرأي العام اليمني

لشهور طويلة وكان المكتب الدائم لاتحاد الصحفيين العرب قد قرر تشكيل لجنة لتقصي الحقائق المتعلقة بشأن حكم الجلد. [25]

Transliteration:

/w>kd AlmsEwdy rfD nqAbp AlSHfyyn llHkm bAEtbArh sAbqp jdydp fy qDAyA Aln$r wTAlb bwDE Hd nhA}y lh bEd >n $gl Alr>y AlEAm Alymny l$hwr Twylpþþ wkAn Almktb AldA}m lAtHAd AlSHfyyn AlErb qd qrr t$kyl ljnp ltqSy AlHqA}q AlmtElqp b$>n Hkm Aljld/

Translation

Al-Massoudi confirmed that the Syndicate of the Journalists refused the judgment being the first in the history of press. He asked for putting an end for such a judgment that was the main concern of the Yemeni public opinion for months. The Permanent Bureau of the Arab Journalists' Union started a fact-finding mission to investigate the judgment.

Some articles, however, show a rather more frequent use of punctuation markers, especially the commas. These articles usually belong to such subject classes as economy, politics and sport, where commas are used to mark phrase and clause boundaries. However, the use of such punctuation markers is not consistent even within the articles belonging to the same genre. Paragraphs (7.2) and (7.3) are extracted from two different economic articles, yet they show inconsistency in using punctuation markers; the commas and the full stops are used in (7.2) to mark every phrase and sentence boundary, whereas in (7.3) only a final full stop is used.

[25] *Al-Ahram Newspaper* Issue No. 40779

(7.2) استعرضت اللجنة الوزارية لمتابعة المشروعات الصناعية الكبري، خلال اجتماعها أمس بحضور الدكتور حمدي البنبي وزير البترول، والدكتور محمد الغمراوي وزير الدولة للإنتاج الحربي ، البرامج الزمنية لتنفيذ المشروعات الصناعية، وشركات التنمية الاستثمارية في كل المنطقة الاقتصادية الخاصة بشمال غرب خليج السويس، حيث تبين وجود بدايات جادة للخطط والبرامج الزمنية للمشروعات.[26]

Transliteration:

/AstErDt Alljnp AlwzAryp lmtAbEp Alm$rwEAt AlSnAEyp Alkbryþ,þ xlAl AjtmAEhA >ms bHDwr Aldktwr Hmdy Albnby wzyr Albtrwlþ,þ wAldktwr mHmd AlgmrAwy wzyr Aldwlp ll<ntAj AlHrby þ,þ AlbrAmj Alzmnyp ltnfy Alm$rwEAt AlSnAEypþ,þ w$rkAt Altnmyp AlAstvmAryp fy kl AlmnTqp AlAqtSAdyp AlxASp b$mAl grb xlyj Alswysþ,þþ Hyv tbyn wjwd bdAyAt jAdp llxTT wAlbrAmj Alzmnyp llm$rwEAtþ/*

Translation

The follow-up ministerial committee of major projects discussed yesterday in a meeting with Dr. Hamdy Albinby, the Minister of Petroleum, and Dr. Mohammed Elghamrawy, the Minister of Military Production, the schedules of the industrial projects, and the investment companies in the industrial zone in the north-west of Suez Cannel. There are serious schemes and schedules.

(7.3) وضع ضوابط علي التدفقات المالية الداخلة والخارجة من والي مصر والتريث في تداول أسهم البورصة المصرية في الأسواق الخارجية وعدم طرح الجنيه المصري ـ في الوقت الحاضر للتعامل الحر في السوق العالمية

[26] *Al-Ahram Newspaper* Issue No. 40779

ومن بين تلك الضوابط فرض ضريبة علي التعاملات قصيرة الأجل في البورصة.

Transliteration:

/wDE DwAbT Ely AltdfqAt AlmAlyp AldAxlp wAlxArjp mn wAly mSr wAltryv fy tdAwl >shm AlbwrSp AlmSryp fy Al>swAq AlxArjyp wEdm TrH Aljnyh AlmSry fy Alwqt AlHADr lltEAml AlHr fy Alswq AlEAlmypþ wmn byn tlk AlDwAbT frD Drybp Ely AltEAmlAt qSyrp Al>jl fy AlbwrSp/

Translation:

Controlling internal and external influxes and rationalizing the exchange of the Egyptian stock market shares in external markets and the blocking the Egyptian pound from free transactions in the international markets ... one of these regulations is the tax of the short-term stock market transactions.

As a result of the inconsistency in using the punctuation markers, they are removed from the corpus even the elongation mark as in بعـــــيد /*bEyd*/ (far) which a specific punctuation marker for Arabic writings used for text highlight and justification and it does not have any syntactic, morphological or semantic indications (Habash 2005).

7.3.2 Numbers

Arabic cardinal numbers are usually followed by التمييز /*Altmyyz*/ ("the specification of number"), which is always an NP following the number (Wright 1981: 124). Such specification can be either singular or plural. Usually, numbers ≥ 10 have a specification that is singular in form but plural in meaning (Hasan 1999). Such

a type of specifications is, therefore, referred to using plural pronouns as in (7.4) where the 3rd person plural masculine pronoun هم /*hm*/ (them) refers to the singular word شخص /*$xS*/.

(7.4) قتل 60 شخصاً من بينهم أطفال

Transliteration:

/qtl 60 $xSA mn bynhm >TfAl/

Translation:

60 persons were killed out of them were some children

If the number is deleted from (7.4), the 3rd person plural pronoun هم /*hm*/ (them) will never be resolved as referring to شخص /*$xS*/ (person) since it is written in the singular form; the plural form being أشخاص /*>$xAS*/ (persons). Thus numbers are the only way to realize that the singular forms of specifications are plural in meaning. As a result, cardinal numbers in the present thesis are part of the corpus tokens. They are normalized to the POS tag of NUM to be counted as one word token and type.

It is unnecessary to know the exact value of the number as far as AR is concerned. If the number equals one, it is unlikely to be found before its specification. A structure such as واحدة فتاة /*wAhdp ftAp*/ (one girl) is very unlikely in Arabic; instead either فتاة /*ftAp*/ or فتاة واحدة /*ftAp wAhdp*/ is used. If the number equals two, then it is also unlikely to be found before its specification, which must be in the dual form as in فتاتان /*ftAtAn*/ (two girls). If the number is ≥ 10, then it is to be found before its specification, which is plural in meaning even if it has a singular word form. Therefore, the NUM tag stands for numbers ≥ 10.

7.3.3 Out-Of-Vocabulary (OOV) Words

As mentioned earlier in section (6.3), different Arabic POS taggers have different strategies to deal with OOV words. In *Al-Ahram Newspaper* corpus, OOV words can be divided into different categories. First, an OOV word can be a misprinted word. For instance, متي /*mty*/ (when) is frequently written with a final *ya'* in *Al-Ahram Newspaper*, although it is correctly written with *alef maqsura* as متى /*mtY*/ (when).

The second category of OOV words includes proper nouns – both Arabic and foreign. Some nouns might not exist in the training set such as some Arabic names and some of the names of foreign politicians, researchers and so on. For example, Diab et al.'s (2004) POS tagger is incapable of tagging the foreign proper noun وايزمان /*wAyzmAn*/ (Wiseman) and the Arabic proper noun عصمت /*ESmt*/ (Esmat: A MASC/FEM Arabic proper noun).

Moreover, OOV words can also be out of the language variety which the POS tagger deals with. Diab et al's (2004) POS tagger deals with MSA; therefore, slang words – which are the third category of the OOV words – such as ماكنش /*mAkn$*/ (it was not) and الواد /*AlwAd*/ (the boy) are tagged as "NOFUN" (i.e. NO FUNCTION).

According to Diab et al.'s (2004) POS tagger – which is the tagger used for POS tagging in the present thesis – ≈ *0.5%* of the corpus is tagged as NOFUN. Being relatively small and being possible antecedents – as in (7.5) below where كابو /*kAbw*/ (CAABU) is a foreign proper noun of an African organization, OOV words are not removed from the corpus.

(7.5) أقول إن كابو رغم أنها اختبار صغير فإنها اختبار دقيق

Transliteration:

/>qwl <n kAbw rgm >nhA AxtbAr Sgyr f>nhA AxtbAr dqyq/

Translation:

I say that CAABU, in spite of being a small test, it is a critical one

7.3.4 Diacritics

As mentioned in section (5.5), diacritics are not usually a part of written MSA (Badawi 1973), which is the main language variety of *Al-Ahram Newspaper*. Thus even if there are few examples of diacriticized words in the corpus, those are deleted.

7.3.5 The Level of Word Analysis

As mentioned earlier in section (5.6), there are two main levels of word analysis chosen in ANLP: the root and the stem. The stem level of analysis is the one chosen in the present thesis because it includes basic information about gender and number which is indispensable for AR. For example, the stem أولاد /*>wlAd*/ (women) shows that it is a masculine, plural noun. Such information is not available for the root ولد /*wld*/ which might be interpreted as a verb (he was born), or a noun (a boy).

7.3.6 Normalization

According to section (5.7), two letters are usually normalized: *alef* and *ya'*. Seemingly, it is an orthographic convention in *Al-Ahram Newspaper* to end all words with *ya'* even if they are originally written with *alef maqsura* as in حتي /*Hty*/ (until), which is correctly written as حتى /*HtY*/. In the entire, corpus no examples of words ending in *alef maqsura* are found. Thus *ya'* is kept as the default in the current tokenization scheme.

As for *alef*, the corpus is inconsistent. For example, the word أمريكا />*mrykA*/ (America) with *a hamza above the alef* is found 183 times and امريكا /*AmrykA*/ (America) with *bare alef* is found 208 times. Such inconsistency leads to normalizing all *alef* variations into *bare alef*.; *alef* without any *hamza*.

7.3.7 Concatenation

According to the definition of concatenation mentioned in section (5.8), three frequent categories of concatenated forms are found in *Al-Ahram*: concatenated function words, concatenated proper nouns and misspelling concatenations.

Buckwalter (2004) states that the two negative particles ما /*mA*/ and لا /*lA*/ tend to concatenate with the following word. In *Al-Ahram Newspaper,* 70% of such words are written in a spaced form. Moreover, some irregular occurrences – 3% – are encountered such as لايرتاب /*lAyrtAb*/ (does not doubt) and مايجري /*mAygry*/ (what is going on).

The Word	The Form	Time of Occurrences
مازال	Concatenated	1
لا يزال	Spaced	2
لايزال	Concatenated	1
لا تزال	Spaced	1
لايرتاب	Concatenated	1
لابد	Concatenated	1
لاشك	——	No instances found

Table (7.1): Concatenations of la and ma

The problem posed by writing these words in two different ways is that spaced words will be handled as two separate tokens/types, and the concatenated form as a third different token/type. For the same reasons mentioned at the beginning of this

chapter, token-type ratio is out of the concerns of the present thesis, thus spaced and concatenated forms of *la* and *ma* are left as they are.

As for concatenated proper nouns, Buckwalter (2004) notices that names starting with عبد /*Abd*/ tend to concatenate with the following name as in عبدالعظيم /*AbdAlEZym*/ (A male masculine name in Arabic). In Arabic, these are compound names that consist of the word *Abd* (the servant of) and one of God's holy names in Islam. Another category of concatenated proper nouns found in *Al-Ahram Newspaper* includes names starting with أبو /*>bw*/ (Abu) or أم /*>m*/ (Om) that also tend to concatenate with the following name as in أبوهريرة /*>bw hryrp/* and أم كلثوم /*>mklvwm*/. In *Al-Ahram Newspaper,* 80% of such proper nouns are written in the concatenated form. Thus they are concatenated to one another.

Concatenating these proper nouns reduces ambiguity by decreasing the search space of the AR algorithm. For example, writing عبد المجيد /*Ebd Almjyd*/ in (7.6) below in the spaced form results in having two possible NP candidate antecedents for the pronoun ه /h/ (he/his): عبد /*Ebd*/ and المجيد /*Almjyd*/, which is very misleading because عبدالمجيد /*EbdAlmjyd*/ refers to one single person, i.e. one entity.

(7.6) وربط الحادثة باتهامات لعبد المجيد تتحدث عن تورطه فيها

Transliteration:

/wrbT AlHAdvp bAthAmAt lEbd Almjyd ttHdv En twrTh fyhA/

Translation:

And he connected the accident to accusations of Abd Al-Megid talking about his involvement in them

Therefore, proper nouns starting with *Abd*, *Abu* or *Om* are written in the concatenated form. God's holy names in Islam are 99; thus controlling them is rather

easy. Two more names can be added since they are frequently used in the Arabic-speaking world: عبد النبي /*EbdAlnby*/ and عبد الرسول /*EbdAlrswl*/.

The last category of concatenated words, especially found in *Al-Ahram Newspaper*, is that of misspelled or misprinted concatenated forms as in قريباًسلامة /*qrybAslAmp*/ (soonSalamh), which is supposed to be written as قريباً سلامة / *qrybA slAmp/* (soon Salamh). Being too few – ≈ 0.03 % of the OOV words – these misspellings are left as they are.

7.3.8 Code-Switching

Due to the intersection between MSA and slang (see section 5.9), slang words cannot be distinguished from MSA words. However, slang is relatively small. In ≈ 4,000 sentences, two slang words have occurred, with ≈ 0.05% probability of occurrence. Thus slang words are not to be removed from the corpus.

7.3.9 Summary

The tokenization scheme used in the present thesis has two purposes to serve: minimizing ambiguity and minimizing human intervention. The main features of such a scheme are summarized in table (4) below.

Tokenization Scheme Task	Decisions Made
Punctuation markers	Removed from the corpus being inconsistently used
Numbers	Normalized to one single tag "NUM" since they are indicative for significations plural in meaning but singular in form
OOV Words	Kept in the corpus being already few and being possible candidate antecedents
Diacritics	Removed for being inconsistently used
The Level of Word Analysis	The stem being less ambiguous than the root
Normalization	*Alef* is normalized to *bare alef* *Ya'* is kept as it is
Concatenation	Proper nouns starting with *Abd, Abu* and *Om* are written in the concatenated form Concatenated function words are split
Code-switching	Slang words are part of the corpus

Table (7.2): The Tokenization Scheme of the Present Thesis

7.4 Corpus Preprocessing Tasks

As mentioned earlier in chapter 6, AR preprocessing tasks typically include non-pleonastic pronoun identification, named-entity recognition, morphological analysis, POS tagging and NP identification. Some tasks might be added or ignored according to the language being dealt with and to the preprocessing tools available for such a language. Each one of these tasks introduces an error rate and thus contributes to the reduction of the performance rate of AR systems. The following subsections discuss preprocessing tasks of the proposed Arabic AR algorithm.

Since AR is a relatively new area of research in ANLP, some of the following corpus preprocessing tasks are preliminarily developed by the researcher so as to know whether AR requires special preprocessing. However, by the end of this part, it will be proved that AR does not require any special corpus preprocessing; its preprocessing tools can be used for any other NLP application.

7.4.1 Tokenization (Light Stemming)

7.4.1.1 Preliminary Tests

The researcher develops and tests a number of tokenizers to know whether Arabic AR requires special type of tokenization. The following are the different tokenizers together with their F-measure performance rates calculated according to manually evaluated random samples.

7.4.1.1.1 Corpus-Based Tokenizer (CorpTok)

The corpus-based tokenizer strips off clitics provided that the striped wordform exists in the corpus. For example, the word والكتب /*wAlktb*/ (and the books) is a word procliticized to the conjunction و /*w*/ (and) and the definite article ال /*Al*/. The corpus-based tokenization algorithm goes as follows:

1. For the word والكتب /*wAlktb*/ (and the books), if الكتب /*Alktb*/ (the books) exists in the corpus, then strip off و /*w*/ (and);
2. If كتب /*ktb*/ (books) exists in the corpus, then strip off ال /*Al*/ (the).

One main advantage of this tokenizer is that it is totally unsupervised. However, the main problem is sparseness of data: a word might not be tokenized only because it does not exist in the corpus.

According to a manual evaluation of a random sample – 1000-word types – and the F-measure evaluation metric, this is the weighted mean of precision and recall, the tokenizer achieves an F-measure score of ≈ 94%.

$$\textbf{F-measure} = \frac{2*(precision * recall)}{precision + recall}$$

7.4.1.1.2 Corpus-Based/Statistically-Based Tokenizer (CorpStatTok)

This tokenizer is a fine-tuned version of the aforementioned corpus-based tokenizer. Due to the inconsistency of the corpus – because of concatenated word forms and misspellings – and word ambiguity, a new condition is added to the above mentioned algorithm. Thus the algorithm now goes as follows:

1. For the word والكتب /*wAlktb*/ (and the books), if الكتب /*Alktb*/ (the books) exists in the corpus, and its relative frequency *F*, which is heuristically assumed, is > 100, then strip off و /*w*/ (and);
2. If كتب /*ktb*/ (books) exists in the corpus, and its relative frequency *F* > 100, then strip off ال /*Al*/ (the).

Adding a statistical condition raises precision and lowers recall. Moreover, the higher the frequency condition, the better the precision is. According to the same random sample used in evaluating the corpus-based tokenizer, this version achieves an F-measure rate of 94.56%.

7.4.1.1.3 Corpus-Based/Dictionary-Based Tokenizer (CorpDictTok)

This version of the tokenizer has one additional condition: the resulting word must exist in a lexicon. This condition is supposed to eliminate errors caused by the disturbances of the corpus itself, especially a corpus like *Al-Ahram Newspaper* which is full of spelling mistakes and misprints. The lexicon used is that of Buckwalter (2002). In this case, the algorithm is as follows:

1. For the word والكتب /*wAlktb*/ (and the books), if الكتب /*Alktb*/ (the books) exists in the corpus, and its relative frequency $F > 100$, and it exists in the lexicon, then strip off و /*w*/ (and);
2. If كتب /*ktb*/ (books) exists in the corpus, and its relative frequency $F > 100$, and it exists in the lexicon, then strip off ال /*Al*/ (the).

Measured on the same aforementioned corpora, this tokenizer achieves an F-measure performance rate of ≈ *95.4%***.**

7.4.1.2 The Used Tokenizer (SVM-TOK)

Since preliminary tests for tokenization show that AR does not require any special tokenization; it is better to use one which is already available and which has higher reported results. The one chosen is that of Diab et al. (2004), which is chosen for the following reasons:

1. It is a public domain tokenizer, which is trained on a huge number of news articles of different topics (Diab et al. 2004).
2. It is the one with the highest accuracy rate in comparison to other existing POS tagger (see section 6.2)
3. It is available for public use.

According to Diab et al. (2004), the SVM-POS achieves an $F_\beta = 1$ of 99.12%. A sample of Diab et al.'s (2004) tokenizer is:

في موقف ها فرضت ب شكل نهائي ما قررت ه مؤقتاً من فرض رسوم مكافحة اغراق على بعض السلع المصرية فان ذلك سوف يلحق الضرر ب الصناعة الاوروبية ذات ها

7.4.2 POS Tagging

7.4.2.1 Preliminary Tests

For preliminary tests, the researcher uses Buckwalter's AraMorph (2002). As mentioned in section (6.3.1.2), Buckwalter's (2002) AraMorph does not include any POS disambiguation procedure; it gives all the possible analyses of the input word, which might be up to 10 analyses due the rich morphological system of Arabic. As a result, 51% of *Al-Ahram* corpus tokens are given more than one POS tag and thus they are marked as ambiguous. In order to disambiguate the output of AraMorph (Buckwalter 2002), the researcher develops the following procedures used in the following sequence:

1. Using the most frequent sense in Buckwalter's lexicon (2002).
2. Using the researcher's corpus-based/dictionary-based tokenizer
3. Using a set of heuristics

Buckwalter (2002) arranges his lexicon according to the most frequent sense of the word in his training corpora. Using the most frequent sense in Buckwalter (2002) achieves a precision rate of 84% manually evaluated on ambiguous words only. Moreover, using the corpus-based/dictionary-based tokenizer, as the second step after using Buckwalter's (2002) most frequent sense, reduces the percentage of ambiguous words to 34%. Finally, a number of heuristics is used.

The heuristics used rely on Arabic function words. The heuristics are summarized in table (7.3) and they achieve a performance rate of **69%**; in spite of being highly accurate, their recall rate is not high.

Arabic Function Word	Function Word Type	The Heuristic
Noun-Disambiguating Cues (NDC)		
من /*mn*/ (from)	Preposition	Any word after a preposition is a NOUN
إلى /*>lY*/ (to)	Preposition	
على /*ElY*/ (on)	Preposition	
مع /*mE*/ (with)	Preposition	
هذا /*h*A*/ (this) ذلك /**lk*/ (that)	Demonstrative	Most of the words after a demonstrative are a NOUN
هذه /*h*h*/ (this) تلك /*tlk*/ (that)	Demonstrative	
هذان /*h*An*/ (these) هذين /*h*yn*/ (these)	Demonstrative	
هاتان /*hAtAn*/ (these) هاتين /*hAtyn*/ (these)	Demonstrative	
هؤلاء /*handlA'*/ (these)	Demonstrative	
أولئك /*>wl}k*/ (those)	Demonstrative	
Noun/Verb-Disambugating Cues (NVDC)		
الذي /*Al*y*/ (who/which)	Relative Pronoun	Any word after a relative pronoun is a VERB Any word before a relative pronoun is a NOUN or an ADJECTIVE
التي /*Alty*/ (who/which)	Relative Pronoun	
اللذان /*All*An*/ (who/which) اللذين /*All*yn*/ (who/which)	Relative Pronoun	
اللتان /*AlltAn*/ (who/which) اللتين /*Alltyn*/ (who/which)	Relative Pronoun	
الذين /*Al*yn*/ (who/which)	Relative Pronoun	
Verb-Disambuagating Cues (VDC)		
لا /*lA*/ (no)	Negative Particle	Any word after a negative particle is a VERB
لن /*ln*/ (no)	Negative Particle	
لماذا /*lmA*A*/ (why)	Interrogative Particle	Any word after an interrogative particle is a VERB
كيف /*kyf*/ (how)	Interrogative Particle	

Table (7.3): Arabic Cues Used for POS Disambiguation

The three aforementioned POS disambiguation procedures result in an F-measure rate of disambiguation of **95%**.

7.4.2.2 The Used POS Tagger (SVM-POS)

The used POS tagger is Diab et al.'s (2004) SVM-POS tagger. The reasons for choosing such a tagger are the same as the reasons for choosing their tokenizer.

According to Diab et al., The SVM-POS tagger achieves an accuracy rate of 95.49%. A sample of the SVM-POS output is:

في/IN زيارة/NN ه/PRP$ هذه/DT التقى/NN إسماعيل/NNP سراج/NNP

ب/IN يلتقي/VBP أن/IN استطاع/NN من/IN كل/IN ب/NNP الدين/NNP

هم/PRP من/IN المفكرين /NNS

where

IN = a preposition, PRP$ = an enclitic pronoun, DT = a demonstrative, NN = a common noun, NNP = a proper noun, VBP = a verb, PRP = a separate pronoun, and NNS = a plural noun.

7.4.3 Semantic Features Acquisition (SFA)

The main motivations for considering SFA as a part of the current methodology are:

1. The semantic features of gender, number and rationality are among the main differences between Arabic and English pronominal systems that cause the poor performance of some current MT systems (see section 3.5).
2. Such features are among the most commonly used semantic constraints in the literature of AR (see section 6.4).

Due to the poor coverage of current Arabic POS taggers and morphological analyzers for such semantic features (6.3.1); the researcher develops an approach to bootstrap them monolingually and bilingually, using the Web as corpus.

7.4.3.1 Monolingual Bootstrapping of SFA

Monolingual bootstrapping of SFA is a cue-based algorithm that depends on cues extracted from the target language itself (i.e. Arabic). The algorithm makes use of the Arabic cues of gender, number and rationality.

The first monolingual bootstrapping seed is extracted from AraMorph's (Buckwalter 2002) output. As mentioned in section (6.3.1.2), AraMorph tags semantic features only when they are morphologically marked. As a result, only 32.8% of the nouns in *Al-Ahram Newspaper* corpus (≈ 20,000,000 tokens; ≈ 971,000 types) are marked for number, 35.5% are marked for gender and 0% is marked for rationality. Such tagged words form the first seed for the monolingual SFA bootstrapping.

The second monolingual bootstrapping seed is built using a set of Arabic number and/or gender cues, which are illustrated in table (7.4).

Arabic Cue	Cue Type	The Features it indicates
ة /*p*/	Suffix	Encliticized to Singular; Feminine Nouns
ون /*wn*/	Suffix	Encliticized to Plural; Masculine; +HUMAN Nouns
ات /*At*/	Suffix	Encliticized to Plural; Feminine Nouns
هذا /*h*A*/ (this) ذلك /**lk*/ (that)	Demonstrative	Followed by Singular; Masculine Nouns
هذه /*h*h*/ (this) تلك /*tlk*/ (that)	Demonstrative	Followed by Singular; Plural Feminine Nouns
هذان /*h*An*/ (these) هذين /*h*yn*/ (these)	Demonstrative	Followed by Dual; Masculine Nouns
هاتان /*hAtAn*/ (these) هاتين /*hAtyn*/ (these)	Demonstrative	Followed by Dual; Feminine Nouns
هؤلاء /*handlA'*/ (these)	Demonstrative	Followed by Plural; Masculine; Feminine Nouns
أولئك /*>wl}k*/ (those)	Demonstrative	Followed by Plural; Masculine Nouns
الذي /*Al*y*/ (who/which)	Relative Pronoun	Preceded by Singular; Masculine Nouns
التي /*Alty*/ (who/which)	Relative Pronoun	Preceded by Singular; Plural; Feminine Nouns
اللذان /*All*An*/ (who/which) اللذين /*All*yn*/ (who/which)	Relative Pronoun	Preceded by Dual; Masculine Nouns
اللتان /*AlltAn*/ (who/which) اللتين /*Alltyn*/ (who/which)	Relative Pronoun	Preceded by Dual; Feminine Nouns
الذين /*Al*yn*/ (who/which)	Relative Pronoun	Preceded by Plural; Masculine; +HUMAN Nouns

Table (7.4): Arabic Cues for the Semantic Features of Gender and Number

The third monolingual seed is built according to the following algorithm:

1. Words encliticized to any of the aforementioned suffixes in table (7.4) are extracted from the corpus.
2. Suffixes are stripped off provided that the resulting word exists in the corpus.
3. Then the resulting word can be tagged for number and gender according to the suffix stripped off.

One example for such an algorithm is the noun الصحفيون /*AlSHfywn*/ (the journalists), given as a result of the algorithm's first step, being encliticized to the plural, masculine and +HUMAN suffix ون /*wn*/. Finding the word الصحفي /*AlSHfy*/ (the journalist) in the corpus runs the second step of the algorithm where the suffix ون /*wn*/ is stripped off. Finally, the word الصحفي /*AlSHfy*/ (the journalist) is tagged as a singular, masculine, +HUMAN noun.

As for rationality, there are two seeds used. The first is a list of proper +HUMAN nouns gathered using Google search engine. The second is a list of verbs which are typically followed by a +HUMAN noun; the list of verbs is given in table (7.5).

The Verb	Its Meaning
ذكر /*kr/	Mention
صرح /SrH/	Declare
أعلن />Eln/	Announce
قال /qAl/	Say
زعم /zEm/	Claim
ناقش /nAq$/	Discuss
قدم /qdm/	Present
أوضح />wDH/	Clarify
عرف /Erf/	Know
وصف /wSf/	Describe
عرض /ErD/	Show
اعتبر /AEtbr/	Consider

Table (7.5): Some Indicating Arabic Verbs for the Rationality Semantic Feature

These monolingual seeds result in a list of ≈ 30,000 tokens/types that is manually filtered.

7.4.3.2 Bilingual Bootstrapping of SFA

Bilingual bootstrapping algorithm is also a cue-based algorithm that uses the cues of one language (i.e. English) to acquire the semantic features of another language (i.e. Arabic).

The bilingual bootstrapping algorithm uses the following tools:

1. English electronic resources: The English WordNet 2.1 (Miller 2005) and English Generic Corpora (Cobb 2004).
2. A set of English cues which are used to search for words with specific semantic features in the aforementioned English resources. All English cues are illustrated in table (7.6).

3. English/Arabic MT systems: Two English/Arabic MT systems are used to guarantee good coverage; the first is Golden Al-Wafi (ATA 2002) and the second is the Google Statistical Machine Translation (SMT) engine[27]

English Cue	Cue Type	Feature(s) it indicates
A/ an/any/ every/ each	Modifier	Followed by Singular Nouns
Some/ all/ any/ many	Modifier	Followed by Plural Nouns
Who	Relative Pronoun	Preceded by +HUMAN Nouns
Which	Relative Pronoun	Preceded by –HUMAN Nouns

Table (7.6): English Cues for the Semantic Features of Number and Rationality

The bilingual bootstrapping algorithm goes as such:

1. The English cues illustrated in table (7.6) are used to extract words from generic English corpora. Moreover, words tagged as ±HUMAN, plural or singular in the English WordNet 2.1 are also compiled.
2. The resulting English words are submitted to Golden Al-Wafi English- and Google SMT engine.
3. Number and rationality semantic features are added to the Arabic noun translations of the English nouns.

For example, the word 'motive' in '… the motive which led this family to …' is extracted from the aforementioned English resources. Since the word precedes the relative pronoun 'which', it is tagged as –HUMAN. Submitted to Golden Al-Wafi and Google SMT engine, the word is translated as دافع /*dAfE*/ which is thus tagged as a –human Arabic noun.

The output list of the bilingual bootstrapping is manually filtered, resulting in a noun-base of ≈ 24,000 types tagged for number and rationality.

[27] The reader is referred to http://www.google.co.uk/language_tools

7.4.3.3 Final Results of SFA

The monolingual and the bilingual algorithms yield the following results:

NUMBER		GENDER		RATIONALITY	
SINGULAR	**PLURAL**	**FEMININE**	**MASCULINE**	**+HUMANN**	**-HUMAN**
26,805	7,083	16, 490	18, 344	4,021	20,477

Table (7.7): Final results of the monolingual and the bilingual algorithms of SFA

These final lists achieve a coverage rate of ≈ 59% for *Al-Ahram* SVM-POS tagged corpus.

7.4.4 Non-Pleonastic Pronouns Identification

According to Badawi et al. (2004), Arabic recognizes non-pleonastic pronouns. Thus one preprocessing step should deal with such pronouns so as to exclude them from the AR algorithm input. As mentioned earlier in section (6.5), in many cases there are no regular patterns that make them practical for a CL algorithm. The only regular patterns of the non-pleonastic pronouns are the ones in table (7.8).

No.	Pattern	Preferred Pronoun's Deletion	Example
1	Relative Pronoun + Verb + Pronoun	Relative Pronoun + Verb + Ø	نفس الكمية التي استهلكتها ... **Transliteration:** */nfs Alkmyp Alty AsthlkthA/* **Translation:** The same quantity consumed by ...
1.a.	Relative Pronoun + *Negation* + Verb + Pronoun	Relative Pronoun + Negation + Verb + Ø	الحجرة التي لم يعمل فيها بالوزارة ... **Transliteration:** */AlHjrp Alty lm yEml fyhA bAlwzArp/* **Translation:** The room in which he didn't work at the ministry...
1.b.	Relative Pronoun + Verb + *Preposition* + Pronoun	Relative Pronoun + Verb + Preposition + Ø	الحجرة التي يعمل فيها بالوزارة ... **Transliteration:** */AlHjrp Alty yEml fyhA bAlwzArp/* **Translation:** The room in which he works at the ministry.
2	Auxiliary Verb (tm/sytm/ytm) + "verbal noun"[28] + Pronoun	Auxiliary Verb (tm/sytm/ytm) + "verbal noun" + Ø	الاتفاقية التي تم توقيعها ... **Transliteration:** */AlAtfAqyp Alty tm twqyEhA/* **Translation:** The agreement that was signed
2.a.	Auxiliary Verb (tm/sytm/ytm) + Negation + "verbal noun" + Pronoun	Auxiliary Verb (tm/sytm/ytm) + Negation + "verbal noun"+ Ø	الاتفاقية التي لم يتم توقيعها ... **Transliteration:** */AlAtfAqyp Alty lm tm twqyEhA/* **Translation:** The agreement that was not signed
2.b.	Auxiliary Verb (tm/sytm/ytm) + "verbal noun" + Preposition + Pronoun	Auxiliary Verb (tm/sytm/ytm) + "verbal noun"+ Preposition + Ø	باستثمارات 10 مليارات دولار يتم الانتهاء منها ... **Transliteration:** */bAstvmArAt 10 mlyArAt dwlAr ytm AlAnthA' mnhA/* **Translation:** With 10-billion investemnts that ended in ...

Table (7.8): Regular Patterns of Non-Pleonastic Arabic Pronouns

Heuristics based on the aforementioned patterns are formed and tested on the LDC parallel Arabic-English corpora. According to such a corpus, these heuristics represent 16.51% of the tokens of the pronouns and 18.83% of the types of the pronouns.

[28] "Verbal noun" or المصدر */AlmSdr/* is the source of the verb (Badawi et al. 2004: 88). It is usually a noun as in يصل */ysl/* (to arrive) and its verbal noun وصول */wswl/* (arrival).

7.5 Summary

Different preprocessing tasks are tackled in the scope of the current thesis. These are all summarized in table (7.9) below.

The Preprocessing Task	The Used Tool	The Source of the Tool	Reasons for Choosing the Tool	The Results of the Tools
Tokenization (Light Stemming)	SVM-TOK	Diab et al. (2004)	- Freely available - Highest performance rate	F-measure rate of 99.12%
POS Tagging	SVM-POS	Diab et al. (2004)	- Freely available - Highest performance rate	F-measure rate of 95%
Semantic Features Acquisition	Bootstrapping Algorithms	The researcher	- The importance of SFA - The insufficient coverage for current ANLP tools	Covering 59% of the words tokens
Non-Pleonastic Pronouns Identification	Heuristically-Based Algorithm	The researcher	- The need to exclude them from the input - The absence of non-pleonastic pronouns identification from previous Arabic AR algorithms	Excluding 16% of the pronoun

Table (7.9): The Used AR-Related Corpus Preprocessing Tasks

Part Four

Algorithm and Discussion

4.1 Introduction

This part mainly focuses on the AR algorithm, its performance and the future work intended to improve it. The first chapter elaborates on the AR algorithm and its specific features including search space, semantic filtration, collocational association, recency and bands. It also highlights the way used to overcome sparseness of data, which is the inherent problem of SNLP. Besides, the chapter focuses on the use evaluation methodology and its yielded results. Finally, an error analysis report is presented so as to point out the weaknesses of the algorithm and to pave the way for the following chapter.

The second chapter concludes the thesis, summarizing the key issues and results and pointing out basic contributions. Moreover, it outlines future directions to improve the proposed AR algorithm and to handle some of the problems encountered throughout the thesis.

Chapter Eight

The AR Algorithm

8.1 Introduction to the Algorithm

The researcher develops a statistical AR algorithm that makes use of the least possible features and achieves 78% precision rate and 87.64% F-measure score. This type of algorithms is described by Mitkov (1998: 1) as "knowledge-poor", since it requires the least syntactic, lexical and morphological resources. Such a type is appropriate for an Arabic AR algorithm for two reasons: first, it does not require much human intervention and thus it saves both time and effort; second, it fits ANLP given the absence of enough ANLP resources and tools.

The algorithm relies on the collocational association between the carrier of the pronoun and the candidate antecedent(s): the candidate antecedent of the strongest collocational association with the carrier of the pronoun is more likely to be the correct one. Such association is measured according to the Conditional Probability (CP) association measure.

The problem with collocational associations is sparseness of data. To get stable associations, massive corpora are required. Such huge corpora are only available through using the Web as corpus, which does not only provide a solution for sparse data, but also adds a dynamic dimension to the algorithm.

Unlike previous statistical, corpus-based approaches, this algorithm does not use any training corpora. Alternatively, the algorithm dynamically generates the collocations necessary to resolve anaphoric pronouns in the input sentences and searches the Web for their frequencies. The absence of a static training corpus handles the problem of sparse data and guarantees a recall rate of ≈ 99%.

Moreover, the algorithm uses the recency feature which relies on the distance between the carrier of the pronoun and the candidate antecedent(s); the closer the candidate is to the carrier, the more likely to be the correct antecedent. Mitkov (1998)

applies this feature, giving it the name of 'referential distance', to Arabic technical manuals for which the feature achieves a precision rate of 34.4% and a recall rate of 98.9%; that is, an F-measure score of 51%.

Some filtration measures (i.e. constraints) are used to reduce the number of the candidate antecedents. The first is limiting the search space to –20 words and further filter it using bands. The second is a POS filter which selects only nouns as possible candidates. The third is the semantic features of number, gender and rationality which are used to select the candidates that match the number, gender and rationality features of the target pronoun.

To sum up, the proposed algorithm works according to the following steps:

- The corpus is preprocessed using the tokenization scheme discussed in chapter 7
- The output of the tokenization scheme is further preprocessed using the previously mentioned AR-related preprocessing tasks: tokenization (light stemming), POS tagging and non-pleonastic pronouns identification.
- Pronouns are detected and the minus-20-word search space is determined for each pronoun.
- Bigrams consisting of the candidate antecedents and the carrier of the pronoun are compiled.
- Such bigrams are filtered using the semantic features of gender, number and rationality.
- Web counts are acquired for the bigrams that pass the semantic filtration.
- The minus-20-word window size is further subdivided into two bands, out of which the band with the highest score is chosen; the score of the band is counted as the summation of the probabilities of the band's bigrams with the carrier of the pronoun.

- The band with the highest score is further divided into smaller bands, whose score is counted in the same way discussed in the previous step.
- The same procedure is repeated until the algorithm gets a single-word band which is supposed to be the correct one.

The following subsections explain each feature, its motivations and problems. Afterwards, the practical evaluation experiments for each individual feature and for the entire algorithm are detailed.

8.2 The Features of the AR Algorithm

8.2.1 Search Space

The search space is the space where the correct antecedent is likely to be found. Usually, many experiments are done to decide on the suitable search space for each language. For example, two sentences backward are found to be the optimal search space for finding the antecedent of a given pronoun in English (Mitkov 1998, Jurafsky and Martin 2002). However, there are no such experiments for Arabic.

The problem with the Modern Standard Arabic (MSA) corpora – such as *Al-Ahram* – is that they either do not use punctuation markers at all or that they inconsistently use them (see section 5.2). As a result, it is difficult to decide on the search space on the basis of the preceding sentence(s). Instead, the researcher uses the concept of the window size: only a number of preceding words that do not necessarily compose a complete linguistic unit. In order to decide on the suitable window size for Arabic, the researcher conducts different experiments, briefly discussed in section (8.4.1) below.

8.2.2 Semantic Filtration

The only semantic filtration used for the proposed algorithm is the semantic features filtration which means that only nouns – within the minus-20-word search

space – that agree in gender, number and rationality with the target pronoun are selected as possible candidate antecedents. Consequently, although there are five candidate antecedents for the ه /*h*/ pronoun in sentence (8.1) below, only السودان /*AlswdAn*/ (The Sudan) is selected as a possible candidate, being +SINGULAR, +MASCULINE and –HUMAN.

(8.1) ... و تعرف حكومة الخرطوم ان ها ضمن القائمة وان قوي اجنبية عديدة تتربص ل السودان و فصل جنوب ه

Transliteration:

/w tErf Hkwmp AlxrTwm An hA Dmn AlqA}mp wAn qwy Ajnbyp Edydp ttrbS l AlswdAn w fSl jnwb h/

Translation:

… and Al-Khartoum knows that it is included in the list and that many foreign powers are waiting for of the Sudan and its Southern area.

The problem with applying the semantic features filtration is the absence of sufficient Arabic NLP resources that deal with them. Consequently, the researcher develops monolingual and the bilingual bootstrapping algorithms to acquire necessary semantic features for Arabic AR (see section 7.4.3).

8.2.3 Collocational Association

Collocational association depends on the relation between the carrier of the pronoun and the candidate antecedents that pass semantic filtration; the candidate antecedent of the strongest collocational association with the carrier of the pronoun is more likely to be the correct one. In order to measure the collocational association between the carrier of the pronoun and the candidate antecedents, the association measure of the Conditional Probability (CP) is used.

Two reasons motivate using CP. First, it is one of the simplest association measures whose results are straightforwardly interpretable unlike both *T-test* and X^2 scores (see section 2.4.2). Second, it is not biased to rare events unlike Pointwise Mutual Information (PMI) and the log-likelihood ratio which makes errors when computing the probabilities of rare events (see section 2.4.2) (Manning and Schutze 2002)

In order to avoid sparseness of data while measuring the collocational association between the carrier of the pronoun and candidate antecedents, the researcher uses the Web as corpus which indeed poses the problem of estimating the size of Arabic Web documents.

In order to use Web frequencies to calculate the conditional probabilities of the target collocations, the Web size (i.e. the size of the Web documents uploaded to the used search engines) must be estimated. Previous studies (Elghamry et al. 2007, Kilgarriff and Grefenstette 2003) estimate the Web size for many languages such as Arabic, English, Italian, German … etc. using the counts of function words as predictors of the Web size.

The following subsections discuss in detail the researcher's methodology in dealing with each of the aforementioned points.

8.2.3.1 Collocational Association and Estimating Arabic Web Size

In order to use the Web as corpus, the researcher relies on two meta-search engines, namely www.search.com and www.alltheweb.com, which support Arabic search on the Web. The Web size for each of the used search engines is estimated according to Kilgarriff and Grefenstette's (2003) equation:

$$\text{Web size} = \frac{\textit{The size of a known corpus} * \textit{Web frequencies for function words}}{\textit{The frequencies of such function words in the corpus of the known size}}$$

Using function words frequencies in table (8.1) below, the equation results in estimating the size of the Arabic Web documents uploaded to each of the aforementioned search engines as 4,500,000,000 Arabic words.

Function Words Frequencies in *Al-Ahram Newspaper*	**Function Words Frequencies on www.alltheweb.com**	**Function Words Frequencies in www.search.com**
394,030	8,894,438,640	4,293,947,945

Table (8.1): Function Words Frequencies in Al-Ahram Corpus and the Two Used Search Engines

8.2.3.2 Collocational Association and Conditional Probability

Estimating the size of Arabic Web documents on the used search engines facilitates using Conditional Probability (CP) as a measure of collocational association between the carrier of the pronoun and the candidate antecedents that pass semantic filtration. The algorithm of calculating conditional probabilities goes as follows:

1. Within the minus-20-word search space, identify suitable candidate antecedents that agree in gender, number and rationality with the pronoun.
2. For each of the resulting candidate antecedents, get the Web frequencies for the antecedent and the carrier of the pronoun
3. Use the acquired Web frequencies to calculate the CP of the candidate antecedent and the carrier of the pronoun
4. Choose the antecedent with the highest CP.

Sentence (8.2) below is a walk-through example to explain the algorithm of the collocational association. Given the minus-20-word search space of the target pronoun ها /*hA*/ (its), the result is:

(8.2) ... المسئولين الايطاليين حول الموقف في الشرق الاوسط في ضوء استمرار التوجه الايطالي ب دعم جهود إعادة عملية السلام الي مسار ها[29]

Transliteration:

/Alms&lyn AlAyTAlyyn Hwl Almwqf fy Al$rq AlAwsT fy Dw' AstmrAr Altwgh AlAyTAly b dEm jhwd AEAdp Emlyp AlslAm Aly msAr hA/

Translation:

… the Italian officials concerned with the situation in the Middle East as a part of the Italian continuous support to the peace process...

Applying semantic filtration leads to three candidate antecedents: جهود /*jhwd*/ (efforts), إعادة /*AEAdp*/ (return) and عملية /*Emlyp*/ (process) which are all –MALE words. Web frequencies for each candidate antecedent are given in table (8.2).

The Candidate Antecedent	**Web Frequencies for the Candidate Antecedents Separately**	**Web Frequencies for the Candidate Antecedents Plus the Carrier of the Pronoun in a remote context**
جهود /*jhwd*/ (efforts)	242,030	29
إعادة /<*EAdp*/ (return)	1,370,030	120
عملية /*Emlyp*/ (process)	590,020	1,070

Table (8.2): Web Frequencies for an Example of Collocational Association

The collocational associations between each one of the candidates and the carrier of the pronoun, namely مسار /*msAr*/ (path), measured according to CP, are:

[29] All examples given in part four are tokenized by SVM-TOK previously discussed in section 7.2.1. They are given in the same form they are given to the algorithm.

The Bigram	Its Conditional Probability Measure
P(/*jhwd*/|/*msAr*/)	0.0408648148
P(/*AEAdp*/|/*msAr*/)	0.0321545644
P(/*Emlyp*/|/*msAr*/)	0.0501978880

Table (8.3): An Example of Candidate Bigrams and Collocational Association

According to table (8.3), the candidate antecedent عملية /*Emlyp*/ (process) has the strongest collocational association with the carrier of the pronoun. Thus it is selected as the correct candidate.

Briefly, collocational association relies on the collocational relation between the carrier of the pronoun and the candidate antecedent(s) which is measured using CP and the Web as corpus to overcome sparseness of data. The Web size for the Arabic language is estimated via function words and it is found to be ≈ 4,500,000,000 words.

8.2.4 Bands

The minus-20-word search space is found to be the most suitable search space (see 8.3.3.1.1) for Arabic AR. Using bands is intended to limit the search space from –20, to –10, to –5, to –2 and to –1, respectively, according to the following algorithm:

1. The –20 words are divided into two bands of –10 words each. These bands are not necessarily complete linguistic units.
2. A score is calculated for each minus-10-word band. The score of the band is the summation of the conditional probabilities of the bigrams of the band; each bigram consists of the carrier of the pronoun and a candidate antecedent.
3. The band of the highest score is chosen to the next step as it is further divided into minus-5-word bands.

4. The score of each minus-5-word band. The score of the band is the summation of the conditional probabilities of the bigrams of the band; each bigram consists of the carrier of the pronoun and a candidate antecedent.
5. The band of the highest score is chosen to the next step as it is further divided into 4 bigrams.
6. The score of each bigram is calculated. The score of the band is the summation of the conditional probabilities of the bigrams of the band; each bigram consists of the carrier of the pronoun and a candidate antecedent.
7. The bigram of the highest score is divided into two unigrams and each bigram gets its score according to its Web counts.
8. The bigram of the highest score is supposed to be the correct antecedent.

A walk-through example of the algorithm of the bands is given below:

(8.3) ه علي سؤال حول موقف السلطة الفلسطينية من وصف ديفيد ساترفيلد نائب مساعد وزير الخارجية الامريكي ل الانتفاضة الفلسطينية بان ها

Transliteration:

/h Ely sandAl Hwl mwqf AlslTp AlflsTynyp mn wSf dyvyd sAtrfyld nA}b msAEd wzyr AlxArjyp AlAmryky l AlAntfADp AlflsTynyp bAn hA/

Translation:

... on a question about the attitude of the Palestinian Authority from the description of David Satterfield – deputy of the assistant of American Foreign Minister – to the Palestinian Intifada as …

The first step is to get the minus-10-word bands:

Band1: ه علي سؤال حول موقف السلطة الفلسطينية من وصف ديفيد

Band2: ساترفيلد نائب مساعد وزير الخارجية الامريكي ل الانتفاضة الفلسطينية بان

The second step is to get the score of each band based on the bigrams' probabilities:

Band1: (0.1603977518) ه علي سؤال حول موقف السلطة الفلسطينية من وصف ديفيد

Band2: (0.7934184451) ساترفيلد نائب مساعد وزير الخارجية الامريكي ل الانتفاضة الفلسطينية بان

Since the score of band2 is higher, then it is further subdivided into two bands:

Band3: (0.165681848) ساترفيلد نائب مساعد وزير الخارجية

Band4: (0.6277365971) الامريكي ل الانتفاضة الفلسطينية بان

The score of band4 is higher than the score of band3, thus it is subdivided into bigrams, excluding function words:

Band5: (0.6277365971) الانتفاضة الفلسطينية

8.2.5 Recency

Recency means that the closer the candidate antecedent is to the pronoun and its carrier, the more likely it is to be the correct one. For instance, in (8.4) there are six possible candidate antecedents that semantically agree with the pronoun ه /*h*/: وجود /*wjwd*/ (existence), اطلاق /*ATlAq*/ (shooting), الرصاص /*AlrSAS*/ (the bullets), الهواء /*AlhwA'*/ (the air), شغل /*$gl*/ (attracting) and الثمن /*Alvmn*/ (the price), respectively. However, the last is the closest and it is the correct antecedent.

(8.4) الي امور شخصية لا وجود ل ها و لكن ه اطلاق الرصاص
في الهواء ل شغل الانظار عن الشواغل و القضايا المهمة و ل تجني
الاثار الثمن الفادح الذي تدفع ه

Transliteration:

*/Aly Amwr $xSyp lA wjwd l hA w lkn h ATlAq AlrSAS fy AlhwA' l $gl AlAnZAr En Al$wAgl w AlqDAyA Almhmp w l tgny AlAvAr Alvnm AlfAdH Al*y tdfE h/*

Translation:

… to personal matters that do not exist but it is distracting attention from important issues and getting the results; the high price that is paid …

Therefore, recency is used as a preferential constraint along with the collocational association feature.

8.2.6 Conclusion

The AR algorithm proposed for the present thesis uses a search space of –20 words, collocational association, bands and recency in order to detect the correct antecedent. Due to the inconsistency or the lack of punctuation marks in MSA corpora, the researcher uses a word-based search space filtered using the semantic features of gender, number and rationality. Collocational association between each candidate antecedent and the carrier of the pronoun is measured according to Conditional Probability which is based on Web frequencies. Candidates are further limited using bands and recency; bands delimit search space based on the conditional probabilities of the bigrams of the bands and recency gives preference to the closest antecedent. The following subsections give the details of the used evaluation methodology and the yielded results.

8.3 Evaluation

Although Mitkov (2001) proposes comparative evaluation (i.e. comparing the proposed system with previous ones) as the best evaluation methodology, it is

difficult to perform it hereby because previous approaches are language dependent (Hobbs 1977 as cited in Mitkov 1999, Lappin and Less 1994 among others); that is, they are specifically designed for the English language. Even a comparison with Mitkov's (1998) study on Arabic is impossible due to missing the original training and testing corpora[30]. As a result, the researcher uses the following evaluation methodology and metrics.

8.3.1 Evaluation Methodology

The training-and-testing paradigm cannot be used for the present algorithm because what is being tested is not a stable model, but a model that is constructed according to the instances (i.e. the pronouns) that are to be resolved. In other words, there is no previous anaphorically-annotated model according to which new instances are being resolved. Instead, new instances are processed on the spot and the results are immediately given. Therefore, no training corpus is being used but only a test corpus.

Alternatively, the researcher uses the gold standard evaluation methodology. A manually-annotated gold standard set which consists of 5000 pronoun types (1000 types for each pronoun) is used. In order to be representative, an intersection from the entire corpus (i.e. ≈ 20,000,000 tokens, ≈ 950,000 types and ≈ 545,441 pronoun tokens) is used to build such a set.

8.3.2 Evaluation Metrics

The used evaluation metrics are precision, recall and F-measure. **Precision** is "a measure of the proportion of selected items that the system got right" (Manning and Schütze 2002: 268). It is calculated as follows:

[30] Personal communication with Prof. Ruslan Mitkov, University of Wolverhampton, UK

$$\textbf{Precision} = \frac{\textit{number of correctly resolved events}}{\textit{number of correctly resolved events} + \textit{number of incorrectly resolved events}}$$

Recall is "the proportion of the target items that the system selected" (Manning and Schütze 2002: 268). It is calculated as:

$$\textbf{Recall} = \frac{\textit{number of correctly resolved events} + \textit{number of incorrectly resolved events}}{\textit{total number of events}}$$

F-measure is the weighted mean of precision and recall. It is calculated as:

$$\textbf{F-measure} = \frac{2 \times (\textit{precision} \times \textit{recall})}{\textit{precision} + \textit{recall}}$$

8.3.3 Evaluation Experiments

8.3.3.1 Evaluation Experiments for Each AR Feature

Using the aforementioned gold standard set, the algorithm is run several times to get the effect of adding/removing each of the used AR features: search space, collocational association, semantic features, bands and recency.

8.3.3.1.1 Search Space

The first experiment to determine the suitable search space for Arabic AR starts with a window size of –20 words; the hypothesis is that a minus-20-word window size might be sufficient because previous studies applied to English show that two preceding sentences form a suitable search space; two sentences might be 20 words long (Mitkov 1998, Jurafsky and Martin 2000). According to the gold standard set, the minus-20-word window size covers 88% of the pronouns.

For the second experiment, the window size is expanded to –40 words which raise the coverage rate to 93%. However, such an expansion of the window size decreases the precision rate (see section 8.4.1).

As for the last experiment, the window size is changed to –20 nouns not words. Tested on the same previous gold standard set, coverage increases to 95%. However, like the second experiment, this expansion decreases precision (see section 8.4.1).

The results of the three experiments are summed up in table (8.4) below:

Experiment Number	**Window Size Tested**	**Reasons for Choosing this particular window size**	**Results**
1	– 20 words	A Heuristically-based Choice	88% Coverage
2	– 40 words	A attempt to increase window size coverage	93% Coverage
3	– 20 nouns	A attempt to get more accurate results	95% Coverage
Final Finding	**Using the first proposed window size (– 20 words) achieves the best results because the other two window sizes reduce the precision rate**		

Table (8.4): Window-size Experiments for Arabic AR

8.3.3.1.2 Semantic Filtration

Running the algorithm on the gold standard set without applying the semantic filtration leads to a precision rate of 52% and an F-measure rate of 68%. Using the semantic lists, complied via the monolingual and the bilingual bootstrapping algorithms discussed in section (7.4.3), increases precision to 65.15% and thus raises the F-measure performance to 79%. Consequently, semantic features are granted as an AR-related feature.

8.3.3.1.3 Collocational Association

Collocational association achieves an F-measure rate of 60% measured on the aforementioned gold standard set. This feature is used as an AR-related feature,

because its result outperforms Mitkov's results (1998) given that the present algorithm deals with unrestricted texts, unlike the technical manuals used by Mitkov (1998) which are lexically and structurally restricted.

8.3.3.1.4 Bands

Bands are added to the minus-20-word search space, semantic filtration and collocational association as another AR-related feature. Separately, bands achieve a precision rate of 50% and an F-measure score of 66.66%, both of which seem promising to raise the overall performance of the algorithm. Thus bands are also used for the final AR algorithm.

8.3.3.1.5 Recency

Adding recency to the present AR algorithm increases precision to 73% and F-measure to 84.4%. This feature is used by Mitkov (1998) and applied to Arabic technical manuals for which it achieves a precision rate of 34.4% and a recall rate of 98.9%; that is, an F-measure score of 51%. That is why recency is also used for the final algorithm.

8.3.3.2 Evaluation Experiments for AR Algorithm

All AR-related features are compiled together and the gold standard set is divided into 5 parts. The algorithm is tested on each part separately, and then a mean performance value is measured. Table (8.5) shows the F-measure performance rates at each phase as well as the mean performance value.

Test Size	Performance
1000	83. 2%
2000	83. 5%
3000	84. 87%
4000	84. 98%
5000	85%
Mean Performance	**≈ 84.4**

Table (8.5): Mean Performance of the AR Algorithm

8.3.3.3 Summary

The proposed AR-related features and the final AR algorithm are evaluated against a manually-annotated gold standard set of 5,000 pronouns. The effect of adding and/or removing each AR-related feature is tested separately so as to determine whether to include it as an AR-related feature. As a result of such an evaluation, the minus-20-word window size, collocational association, semantic features, recency and bands are all considered AR-related features.

The same gold standard set is used to evaluate the results of integrating all the abovementioned AR-related features. This yields the following results:

Recall	Precision	F-Measure
100%	78%	84.4%

Table (8.6): Total Evaluation Results of the AR Algorithm

8.4 Error Analysis

Earlier in section (2.4.4), it is mentioned that an *error analysis* indicates how an algorithm may be improved in subsequent redesigns; being able to inspect the types of errors that are being made, and being able to generalize them into linguistic features are useful for the subsequent redesign of the statistical model used by the algorithm (Burch and Osborne 2003).

As for the proposed algorithm, the window size, the output of the POS tagger and Web frequencies are the main sources of errors. The following lines discuss each error and present an overview of the experiments conducted in an attempt to overcome each one of them.

8.4.1 Window Size

According to the evaluation results, 12% of the errors are caused by the insufficient window size; that is, the correct antecedent exists outside the minus-20-word search space. As a result, two experiments are conducted to make up for this inefficiency.

8.4.1.1 Experiment 1 for Window Size Make Up

The researcher first enlarges the window size to −40 words which raises window size coverage to 93% measured according to the same gold standard set mentioned in section (8.3.1). However, precision rate decreases to 67% and thus the F-measure decreases to 80%, compared to the 82% achieved with the minus-20-word window size. The minus-40-word window size is excluded.

8.4.1.2 Experiment 2 for Window Size Make Up

The researcher changes the window size to –20 nouns rather than words. Tested on the previous gold standard set, precision rate also decreases to 69% and the F-measure rate to 81.6%. Consequently, the minus-20-noun window size is excluded.

As a result of the two aforementioned experiments, the minus20-word window size is considered the most suitable search space. In both experiments, precision decreases because the wider the search space is, the more candidates are introduced to the algorithm. Therefore, a band may get a higher score than the other, only because it contains more candidate antecedents, even after applying semantic filtration. This is exemplified in (8.5):

(8.5) انوثة ها ب حجم وافر من الكبرياء كانت مضربة عن الطعام احتجاج علي قانون نقابة المهن الفنية و كانت المرة الاولي التي اراها في ها و لم اكن اعرف ب دقة هل جئت اليها بحث عن رؤية حلم قديمة صنعت ه (Target Pronoun)

Transliteration:

/Anwvp hA b Hjm wAfr mn AlkbryA' kAnt mDrbp En AlTEAm AHtjAj ElY qAnwn nqAbp Almhn Alfnyp wkAnt Almrp AlAwlY Alty ArAhA fy hAw lm Akn AErf b dqp hl j}t AlyhA bHv En r&yp Hlm qdymp SnEt h /

Translation:

Supported by her proud femininity, she started a food strike to object to the law of the Art Syndicate. It was my first time to see her. I did not know whether I came to her in search for an old dream …

The first band in (8.5) contains five candidates: حجم /*Hjm*/ (volume), الكبرياء /*AlkbryA'*/ (the pride), الطعام /*AlTEAm*/ (the food), احتجاج /*AHtjAj*/ (objection) and قانون /*qAnwn*/ (law). The second band contains only two candidates: بحث /*bHv*/ (search) and حلم /*Hlm*/ (dream). That is why the total score of the first band is 0.29 compared to the total score of the second, 0.03. Accordingly, the first band – the one with the highest score – is mistakenly selected as the correct band.

8.4.2 POS Tagging

According to the evaluation results, 5% of the errors are related to POS tagging[31]; that is, 5% of the words are tagged as encliticized to a pronoun, although it is only a part of the word. A sample of such words is illustrated in table (8.7):

[31] The POS tagger used is the SVM-POS tagger of (Diab et al. 2004) (see 7.4.2.2)

The Word	SVM-POS Analysis	Correct Analysis
مياه /*myAh*/	ميا /*myA*/ + ه /*h*/	مياه /*myAh*/ (water)
بنها /*bnhA*/	بن /*bn*/ + ها /*hA*/	بنها /*bnhA*/ (Bnha: An Egyptian City)
جيزه /*jyzh*/	جيز /*jyz*/ + ه /*h*/	جيزه /*jyzh*/ (Giza: An Egyptian Town)
متهما /*mthmA*/	مت /*mt*/ + هما /*hmA*/	متهما /*mthmA*/ (convicted)

Table (8.7): Examples of Words Mistakenly Tagged by the SVM-POS

The researcher proposes a methodology to overcome this 5% error. Typically, if the last part of the word is a pronoun, then the stemmed word must occur in the corpus. The Web is used as a corpus to find the stemmed words: if the stemmed word is found on the Web, then it is considered a word encliticized to a pronoun; otherwise it is ignored. This is explained in detail in table (8.8):

The Word	Web Frequency of Stemmed Word [32]	Web Frequency of Non-Stemmed Word	Decision
مياه /*myAh*/	75,000	1,120,000	The last letter is part of the word
بنها /*bnhA*/	15,700,000	53,400	The last two letters are a pronoun
جيزه /*jyzh*/	12,600	17,900	The last letter is part of the word
متهما /*mthmA*/	494,000	231,000	The last three letters are a pronoun

Table (8.8): Using the Web as corpus to Overcome the Errors of the SVM-POS

This method reduces the 5% error rate to 2% and increases precision to 77.6%. Consequently, the final result for the algorithm is a performance rate of *87.4%*.

8.4.3 Web Frequency Errors

Approximately, 3% of the errors are caused by the inaccuracy of the Web frequencies. For example, in (8.6) below the two bands have an equal number of candidates, five nouns each. The Web frequencies of the candidate antecedents in the first band – علوي /*Elwy*/ (Olwy: a masculine Arabic proper noun) and المكياج /*AlmkyAj*/ (the make-up) – are higher than the Web frequencies of the candidate antecedents –

[32] www.alltheWeb.com is the Website used for such a method

فيلم /*fylm*/ (movie) and نور /*nwr*/ (light) in the second band. As a result, the score of the first band is higher than the second which is the correct one.

(8.6) ل يلى علوى تخلت الفنانة ليلي علوي عن الماكياج من اجل فيلم نور و نار ...

Transliteration:

/l ylY ElwY txlt AlfnAnp lyly Elwy En AlmkyAj mn Ajl fylm nwr w nAr .../

Translation:

The actress Laila Olwy gave up her make-up for the sake of her new movie "Nour w Nar" (Light and Fire) starring...

In short, the major three errors of the proposed algorithm are attributed to the window size, the output of the POS tagger and the Web frequencies. In spite of the relatively high error rate that results from the used minus20-word window size, it is found that this window size is the most convenient one in terms of precision. The error rate of the POS tagger is cut by 3%, using the Web as corpus so as to know whether a given group of letters is a part of a word or an enclitic pronoun. Finally, more work is still required to overcome the 3% error rate caused by Web frequencies.

8.5 Conclusion

The statistical, corpus-based AR algorithm used for the present thesis makes use of five AR-related features – a search space of –20 words, the collocational association between the carrier of the pronoun and each of the possible candidate antecedents, the semantic features of gender, number and rationality, recency and bands. Each one of these features is evaluated against a manually-crafted gold standard set of 5000 types of pronouns in order to examine its performance separately

and/or the effect of adding or removing it. Table (8.9) summarizes the evaluation results for each feature.

The Algorithm's Feature	The Feature's Performance
Search Space	88% coverage
Collocational Relations	60%
Bands	66.66%
Semantic Features	79%
Recency	84.4%

Table (8.9): The Features of the AR Algorithm

When integrated together and evaluated against the abovementioned gold standard set, all features yield the following results:

Recall	Precision	F-Measure
100%	78%	87.4%

Table (8.10): Total Evaluation Results of the AR algorithm

Chapter Nine

Conclusion and Future Work

9.1 Introduction

This chapter first summarizes the present thesis, highlights the key issues being discussed, the contributions being achieved, the methodologies being used and the problems being posed. Second, it elaborates on future directions to improve the proposed AR algorithm.

9.2 A Summary

Anaphora Resolution (AR) is the process of determining the antecedent of a given anaphor (Deoskar 2004, Mitkov 1999, Mitkov 2001). It is required for many NLP applications, such as Question Answering (QA), Information Extraction (IE), Text Summarization (TS), Machine Translation (MT) … etc. (Deoskar 2004, Mitkov 1999). Furthermore, AR is important for intermediate NLP tasks such as Word Sense Disambiguation (WSD), Prepositional Phrase (PP) attachment ... etc. (Mitkov 2001).

Anaphora is typically classified according to the type of the anaphor and the position of the antecedent. There are many types of anaphor, including indefinite NP anaphora, definite NP anaphora, pronominal anaphora, *one* anaphora and demonstratives (Deoskar 2004, Jurafsky & Martin 2000, Mitkov 1999, Werth 1999). The position of the antecedent divides anaphora into intrasentential and intersentential anaphora (Werth 1999). The present thesis basically focuses on pronominal anaphors both intrasentential and intersentential.

The importance of AR is not the only motivation for the present thesis. The thesis is also motivated by the poor performance of some current Machine Translation (MT) systems – such as Sakhr, Google and Systran – in terms of Arabic/English AR. Such poor performance is mainly attributed to the differences between the Arabic and English pronominal systems in terms of gender, number, morphology and grammatical cases as summarized in table (9.1) below:

The Difference	In Arabic	In English
Gender	Two genders: masculine or feminine	Three genders: masculine, feminine or neutral
Number	Five types of number: singular, dual, plural, collective and non-count nouns	Four types of number: singular, plural, collective and non-count nouns
Grammatical Case	Three grammatical cases with a different form each	Three grammatical cases with the same form each
Morphology	Some pronouns are only separate, some are only enclitics and others are both separate and enclitics	Pronouns are typically separate

Table (9.1): Differences between Arabic and English pronominal systems

Another motivation for the present thesis is the fact that AR is an understudied issue in Arabic NLP. To the best of the researcher's knowledge, the only study about Arabic AR is that of Mitkov (1998), which focuses on AR in Arabic technical manuals known for being semantically and syntactically restricted. The fact that AR requires many NLP morphological, semantic and syntactic resources and that Arabic NLP lacks such resources (Diab et al. 2004) makes Arabic AR a difficult issue to handle.

Consequently, the proposed AR algorithm uses a statistical, corpus-based approach which relies on the least resources available and is thus described as 'knowledge poor'. First, it only uses tokenization and POS tagging for corpus preprocessing; both of which are available through Diab et al.'s (2004) SVM package. Second, it requires the least semantic information represented by the semantic features of gender, number and rationality and the collocational association between the carrier

of the pronoun and the candidate antecedent(s). Third, it does not use any syntactic information; it even uses a word-based search space. Besides, the only discourse-based feature used is that of recency, which can easily be depicted.

In other words, none of the used AR-related features is knowledge-rich. The collocational association feature is based on the association between the carrier of the pronoun and the candidate antecedent provided that it is a noun and that it semantically matches the pronoun. Recency gives preference to the close candidates. The semantic features of gender, number and rationality are extracted using monolingual and bilingual semi-automatic algorithms. Finally, the used bands and the search space are only groups of words not complete linguistic units.

This AR algorithm, however, faces two problems: sparseness of data and the lack of Arabic AR-related resources like Arabic semantic features taggers and Arabic non-pleonastic pronouns identifiers. As for sparse data, the researcher follows a linguistically-based approach using the Web as corpus to get the frequencies of the bigrams and thus to measure the Conditional Probability (CP) for each bigram. One main problem with using the Web frequencies is the unknown size of the Arabic Web documents. Such a problem is dealt with using Kilgarriff and Grefenstette's (2003) equation:

$$\text{Web size} = \frac{\textit{The size of known corpus} * \textit{Web frequencies for function words}}{\textit{The frequencies of the function words in the corpus of the known size}}$$

According to this equation and the two used meta-search engines – www.search.com and www.alltheweb.com, the size of the Arabic Web documents is ~ 4,500,000,000 Arabic words. Estimating the Web size facilitates using collocational association and conditional probabilities and avoiding sparseness of data.

The lack of Arabic semantic features taggers and Arabic non-pleonastic pronouns identifiers is another problem for Arabic AR. In order to acquire the necessary semantic features of the Arabic candidate antecedents, the researcher uses monolingual and bilingual bootstrapping algorithms based on Arabic and English cues, respectively. Both algorithms yield to a covering rate of ~ 59% of the nouns in *Al-Ahram* corpus. The problem of the Arabic non-pleonastic pronouns identifier is sorted out using a rule-based algorithm to extract/exclude non-pleonastic pronouns from the AR input. Based on Arabic grammatical rules, the algorithm manages to correctly exclude 16% of the pronouns as non-pleonastic.

Since the proposed algorithm is a dynamic one; that is, there is no stable training model according to which the output of the algorithm is evaluated, the researcher uses the gold standard set methodology for evaluation. This set, which consists of 5,000 pronouns, is manually annotated for anaphoric relations and is used to evaluate each of AR-related features and the entire AR algorithm. Using the evaluation metrics of precision, recall and F-measure, the algorithm achieves a performance rate of 87.4%.

According to the researcher's error analysis, the errors of the AR algorithm are mainly attributed to the limitations of the search space, the output of the POS tagger and the Web frequencies. The minus-20-word window size covers only 88% of the tested pronouns. However, experiments to make up for such insufficiency decreases precision; thus the minus-20-word search space is considered the most optimal size. The POS tagger yields 5% error rate which is decreased to 2%, using the researcher's Web-based tokenizer. Finally, Web frequencies are not necessarily the most accurate.

The proposed AR algorithm has some contributions. First, it is the first algorithm, to the best of the researcher's knowledge, to deal with AR in unrestricted

naturally-occurring Arabic corpora. Second, it makes use of a purely statistical, corpus-based approach, and thus overcomes the bottleneck of scarce Arabic NLP resources and tools. Third, it introduces a new dimension to the problem of scarce Arabic NLP resources and tools through using bilingual resources. Finally, it introduces the Web as corpus as a way to develop an AR algorithm that overcomes the problem of sparse data and guarantees a high recall rate.

9.3 Future Work

The AR algorithm can be improved in various ways that overcome some of its current errors and that add more resources. As for collocational association, more association measures should be used and compared with the measure of the Conditional Probability (CP). In brief, CP is used only as an association measure not as the association measure.

Acquiring more accurate Web frequencies is another possible way to improve performance. Inaccurate Web frequencies are basically related to typos, misprints, spelling and grammatical mistakes and adult content Web pages. In order to overcome these errors, more meta-search engines, like www.exalead.com and www.findforward.com, are to be examined and evaluated for accuracy. Besides, Arabic-based search engines, like www.amamk.com, www.araby.com and www.ayna.com, should be included. The main difference between the two types of search engines is that the Arabic-based search engines use Arabic resources such as dictionaries and morphological analyzers, in order to get as accurate search results as possible[33].

[33] See the homepages of www.araby.com, www.amamk.com and www.ayna.com. Accessed: 2 January 2008

Another way to improve performance is to handle the window size. Instead of expanding the window size to the –40 words or changing it to –20 nouns, chains can be used. The concept of the chains states that each pronoun should be substituted with its antecedent iteratively. For example, the antecedent of the ه /*h*/ pronoun in (8.6) falls outside the minus-20-word window size; it is actually the –38 word.

(8.6) فان باراك قد اسقطت ه الانتفاضة و اهدرت هيبت ه قبل ان تسقط ه الصراعات الحزبية الاسرائيلية بعد فشل ه المطلق في كل سياسات ه و مناورات ه لقد تقطعت خطوط اتصالات ه و اهتزت دفاعات ه و انكشفت مغامرات ه

Transliteration:

/fAn bArAk qd AsqTth h AlAntfADp w Ahdrt hybt h qbl An tsqT h AlsrAEAt Alhzbyp AlAsrA}ylyp bEd f$l h AlmTlq fy kl syAsAt h w mnAwrAt h lqd tqTEt xTwT AtSAlAt h w Ahtzt dfAEAt h w Ank$ft mgAmrAt h/

Translation:

Barrack has been defeated by the Intifada which has destroyed his dignity even before his defeat by the Israeli parties' conflicts that have led to his absolute failure in all his policies and maneuvers. His communications, defenses and adventures have been destroyed.

In (8.6), there are eight ه /*h*/ pronouns before the last one; all of these pronouns refer to باراك /*bArAk*/ (Barrack). Thus if each of the eight pronouns is replaced by the correct antecedent, the antecedent of the target pronoun will be the word –3. The algorithm can informally be described as follows:

1. The first occurrence of the pronoun is substituted by the correct antecedent:

   اسـقطت **بـاراك** الانتفاضـة و اهـدرت هيبـت ه قبـل ان تـسقط ه الـصراعات الحزبية الاسرائيلية بعد فشل ه المطلق فـي كل سياسـات ه و منـاورات ه لقد تقطعت خطوط اتصالات ه و اهتزت دفاعات ه و انكشفت مغامرات ه

2. The second occurrence of the pronoun is substituted by the correct antecedent:

   هيبـة **بـاراك** قبـل ان تـسقط ه الـصراعات الحزبيـة الاسـرائيلية بعـد فـشل ه المطلـق فـي كـل سياسـات ه و منـاورات ه لقد تقطعت خطوط اتـصالات ه و اهتزت دفاعات ه و انكشفت مغامرات ه

3. The third occurrence of the pronoun is substituted by the correct antecedent:

   تـسقط **بـاراك** الـصراعات الحزبيـة الاسـرائيلية بعـد فـشل ه المطلـق فـي كـل سياسات ه و مناورات ه لقد تقطعت خطوط اتـصالات ه و اهتزت دفاعـات ه و انكشفت مغامرات ه

4. The fourth occurrence of the pronoun is substituted by the correct antecedent:

   فـشل **بـاراك** المطلـق فـي كـل سياسـات ه و منـاورات ه لقـد تقطعـت خطـوط اتصالات ه و اهتزت دفاعات ه و انكشفت مغامرات ه

5. The fifth occurrence of the pronoun is substituted by the correct antecedent:

   سياسـات **بـاراك** و منـاورات ه لقـد تقطعـت خطـوط اتـصالات ه و اهتـزت دفاعات ه و انكشفت مغامرات ه

6. The sixth occurrence of the pronoun is substituted by the correct antecedent:

   منـاورات **بـاراك** لقـد تقطعـت خطـوط اتـصالات ه و اهتـزت دفاعـات ه و انكشفت مغامرات ه

7. The seventh occurrence of the pronoun is substituted by the correct antecedent:

   اتصالات **باراك** و اهتزت دفاعات ه و انكشفت مغامرات ه

8. The eighth occurrence of the pronoun is substituted by the correct antecedent:

اهتزت دفاعات **باراك** و انكشفت مغامرات ه

9. The last occurrence of the pronoun is substituted by the correct antecedent:

انكشفت مغامرات **باراك**

Using chains will not run the risk of adding unnecessary candidates and thus more noise to the input of the algorithm. Meanwhile, they are very likely to resolve the problems caused by the insufficient search space.

Another way to improve the performance of the proposed AR algorithm is to develop a system for sentence boundaries identification. As mentioned earlier in (section 5.2), Modern Standard Arabic (MSA) corpora lack consistent punctuation schemes and as a result it is rather difficult to identify sentence boundaries. Identifying sentence boundaries will not only help finding the window size more accurately, but it will also add more heuristics to the AR algorithm which might improve performance. Such heuristics might include the identification of the subject and/or the topic which are successfully used by (Mitkov et al. 1998, Lappin and Leass (1994), Bean and Riloff (2004) among others).

Through evaluation experiments, it is proved that using the semantic features of gender, number and rationality improves the performance of the proposed AR algorithm by ~ 11%. Thus a more comprehensive database of semantic features might also raise efficiency rate of performance. The database can be expanded through adding more features, using more cues and more search engines.

Nominal chunkers might also positively contribute to the performance of the AR algorithm. Currently, the only Arabic chunker that exists is Diab et al.'s (2004) which provides insufficient Arabic chunks as in (8.7):

(8.7) لقد تفكك الاتحاد السوفيتي من داخل ه و ظهرت النزاعات و الزعامات و كل ها ضد النظام القائم الذي لم يحقق ل الشعوب السوفيتية ما كانت تحلم ب ه

Transliteration:

*/lqd tfkk AlAtHAd Alswfyty mn dAxl h w Zhrt AlnzAEAt w AlzEAmAt w kl hA Dd AlnZAm AlqA}m Al*y lm yHqq l Al$Ewb Alswfytyp mA kAnt tHlm b h/*

Translation:

The Soviet Union has been internally destroyed as conflicts and leaderships have appeared against the existing system which has not achieved what the Soviet people have dreamt of

Sentence (8.7) has at least four nominal chunks (i.e. NPs) which are الاتحاد السوفيتي */AlAtHAd Alswfyty/* (The Soviet Union), النزاعات و الزعامات */AlnzAEAt w AlzEAmAt/* (conflicts and leaderships), النظام القائم */AlnZAm AlqA}m/* (the existing system) and الشعوب السوفيتية */Al$Ewb Alswfytyp/* (The Soviet people). However, none of these NPs is recognized by Diab et al. (2004) since it yields only one-word chunks. Using a nominal chunker reduces the search space and the number of candidate antecedents, and gets more accurate association results; instead of getting the probabilities of individual words, the algorithm can work on the probabilities of complete chunks.

Finally, the proposed algorithm is tested on *Al-Ahram Newspaper* corpus. To guarantee its performance, the algorithm should be tested on different corpora types: newswire, literary and annotated corpora.

References

Abdelali, A., Cowie, J. and Soliman, H. (2005). Building a Modern Standard Arabic Corpus. *Proceedings of Workshop on Computational Modeling of Lexical Acquisition*, Fez, Morocco.

Abney, S. (2000). Statistical Methods. In *Encyclopedia of Cognitive Science* (Vol. 1, pp. 283-289). London: Macmillan Ltd.

Albehairy, S. (2003). *Modkhal Ela Elm Alnass: Moshkelat Bena'a Alnass* (An Introduction to Text Linguistics: Problems of Text Building). Cairo: Moassah Al-Mukhtar lil Nasher w Al-teba'a.

Alfiky, S. (2000). *Elm Al-Lugha Al-Nasi Baiyn Al-Nazari'a w Al-tatbeeq* (Text Linguistics: Theory and Practice: A Case Study of Mecca Suras). Cairo: Daar Qiba'a.

Algilayyeny, M. (2003). *JameE Aldruws AlErabyp* (Arabic Lessons Collection). Lebanon: Almaktebh Alesri'a.

Al-Hafez, M., Clarke, M., and Vella, A. (1994). A Semantic Knowledge-Based Computational Dictionary. *Machine Translation: Ten years on Proceedings of the Second International Conference*, UK, 1-13.

Alhashemy, A. (2000). *Al-Quawaed Al-Asasiy'a Lillugha Al-Arabia* (Basic Rules of the Arabic Language). Lebanon: Daar Al-Kitab Al-Alamyia.

Ali, N. (1993). Machine Translation: A Contrastive Linguistic Perspective [Online]. Available: http://www.unesco.org/comnat/france/ali.htm Accessed: 18 April 2007

Aone, C. and McKee, D. (1993). A Language-Independent Anaphora Resolution System for Understanding Multilingual Texts. *Proceedings 31st Annual*

Meeting of the Assoc. for Computational Linguistics (ACL-93), Columbus, Ohio, 156-163.

ATA Software Technology Ltd. (2002). Golden Al-Wafi Translator Software. Version 1.12 [Online]. Available: www.**ata**soft.com Accessed: 21 April 2001.

Badawi, S. (1973). *MstwyAt AlErbyp AlmuEASirp fy Misr* (Modern Arabic Levels in Egypt). Cairo: Daar Almaeref

Badawi, S, Michael, G, and Gully, A. (2004). *Modern Written Arabic: A Comprehensive Grammar*. London; New York: Routledge.

Baldwin, B. (1997). CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Research. *Proceedings of the Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, Madrid, Spain, 38-45.

Bean, D. and Riloff, E. (2004). Unsupervised Learning of Contextual Role Knowledge for Coreference Resolution. *Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL-04)*, May 2004, Boston, USA, 297-304.

Beesley, K. (2001). Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. *Proceedings of ACL/EACL 2001*, Toulouse, France, 1-8.

Berk, L. (1999). *English Syntax: From Word to Discourse.* Oxford; New York: Oxford University Press.

Blum, A. and Mitchell, T. (1998). Combining Labeled and Unlabeled Data with Co-training. *Proceedings of the Workshop on Computational Learning Theory*. Wisconsin, USA, 92-100.

Bond, F., Ogura, K. and Kawaoka, T. (1995). Noun Phrase Reference in Japanese-to-English Machine Translation. *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI '95)*, Leuven, Belgium, 1-14.

Brashi, A. (2005). *Arabic Collocations: Implications for Translation.* Unpublished PhD, University of Western Sydney, Australia.

Brown, P., Pietra, S. and Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19, 263-311

Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. LDC. Catalog number LDC2002L49, ISBN 1-58563-257-0.

Buckwalter, T. (2004). Issues in Arabic Orthography and Morphology Analysis. *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (COLING 20004)*, 31-34.

Burch, C. and Osborne, M. (2003). Statistical Natural Language Processing. In Farghaly, A. (Ed.), *A Handbook for Language Engineers* (pp. 1-31). USA: Center for the Study of Language and Information.

Bussmann, H. (1996). *Routledge Dictionary of Language and Linguistics.* London; New York: Routledge.

Carbonell, J. and Brown, R. (1988). Anaphora Resolution: A Multi-Strategy Approach. *Proceeding of the 12th International Conference on Computational Linguistics (COLING'88)*, Hungary, 96-101.

Chalabi, A. (2001). Sakhr Web-Based Arabic-English MT Engine. *Proceedings of the Association for Machine Translation in the Americas (AMTA'98)*, Toulouse, France, 518-521.

Chapelle, O., Schölkopf, B. and Zein, A. (2006). *Semi-Supervised Learning*. USA: MIT Press.

Chowdhury, G. (2003). Natural Language Processing. *Annual Review of Information Science and Technology*, 37, 51-89.

Cobb, T. (2004). Web Concordancer [Online]. Version 4.0. Available: www.lextutor.ca Accessed: Nov. 17, 2006.

Darwish, K. and Oard, D. (2002). CLIR Experiments at Maryland for TREC-2002: Evidence Combination for Arabic-English Retrieval. *Proceedings of the 11th Text Retrieval Conference (TREC 2002)*, National Institute of Standards and Technology (NIST).

Dagan, I. and Itai, A. (1990). Automatic Processing of Large Corpora for the Resolution of Anaphora References. *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, Helsinki, Finland, 330-332.

Deoskar, T. (2004). Techniques for Anaphora Resolution: A Survey [Online] Available: www.cs.carnell.edu Accessed 12/9/2006.

Diab, M., Hacioglu, K. and Jurafsky, D. (2004). Automatic Tagging of Arabic Text: from Raw Text to Base Phrase Chunks. In Dumas, S., Marcus, D. and Roukos, S. (Eds.). *HLT-NAACL 2004: Short Papers* (pp.140-152). Boston: Association for Computational Linguistics.

Elghamry, K, El-Zeiny, N. and Al-Sabbagh, R. (2007). Arabic Anaphora Resolution Using the Web as Corpus. *Proceedings of the Seventh Conference of Language Engineering, The Egyptian Society of Language Engineering*, December 2007, Cairo, Egypt

Elhaddad, N. (2006). Natural Language Processing [Online] Available: http://www-cs.engr.ccny.cuny.edu Accessed: 19 July 2007.

Evert, S. and Krenn, B. (2003). Computational Approaches to Collocations [Online] Available: http://www.collocations.de Accessed: 24 July 2007.

Freeman, A. (2001). Brill's POS Tagger and a Morphology Parser for Arabic. *Proceedings of the 39th Annual Meeting of Association for Computational Linguistics and 10th Conference of the European Chapter, Workshop on Arabic Language Processing: Status and Prospects*, Toulouse, France.

Fujii, A. and Ishikawa, T. (2000). Utilizing the World Wide Web as an Encyclopedia: Extracting Term Descriptions from Semi-Structured Text. *Proceedings of 38th Meeting of the Association for Computational Linguistics*, Hong Kong, 488-495.

Gaber. M. (1980). *Al-Damayer fy Allugha AlArabia* (Pronouns in the Arabic Language). Cairo: Daar Al-Maeref.

Ge, N., Hale, J. and Charniak, E. (1998). A Statistical Approach to Anaphora Resolution. *Proceedings of the 6th Workshop on Very-large Corpora*, Montréal, Canada, 161-170.

Glossary. (2008). [Online]: http://www.rand.org/pubs/monograph_reports/MR1005/MR1005.gloss.pdf. Accessed: 10 April 2008

Grefenstette, G., Semmar, N. and Elkateb, F. (2005). Modifying a Natural Language Processing System for European Languages to Treat Arabic Information Processing and Information Retrieval Applications. *Proceedings of the ACL Computational approaches to Semitic Languages*, USA, 31-38.

Habash, N. (2004). Large Scale Lexeme Based Arabic Morphological Generation. *Proceedings of Session Traitement Automatique de l'Arabe* (*JEP-TALN 2004*), Fez, Morocco.

Habash, N. (2005). Introduction to Arabic Natural Language Processing. *Proceedings of ACL'05.* A Three-hour tutorial in Johns Hopkins University Summer School on Human Language Technology, Baltimore.

Habash, N. (2007). Syntactic Preprocessing for Statistical Machine Translation. *Proceedings of the Machine Translation Summit (MT-Summit),* Copenhagen, Denmark, 2007.

Habash, N. and Rambow, O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Michigan, USA, 573-580.

Halliday, M. (1985). *An Introduction to Functional Grammar*. London: Edward Arnold Publishers Ltd.

Halliday, M. and Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hasan, A. (1999). *AlnaHw AlwAfy* (The Comprehensive Grammar). 4 vols. Cairo: Daar Almaeref.

Hutchins, J. (2005). Towards a Definition of Example-Based Machine Translation. *Proceedings of the Second Workshop on Example-Based Machine Translation*, Phuket, Thailand.

Ivanova, P. (2000). Introduction to Computational Linguistics. [Online] Available: www.ssw.uni-linz.ac.at Accessed: 17 July 2007.

Jones, R. (2005). *Learning to Extract Entities from Labeled and Unlabeled Texts*. PhD Thesis from Carnegie Mellon University, School of Computer Science. Pittsburgh: USA.

Jones, R. and Ghani, R. (2000). Automatically Building a Corpus for a Minority Language from the Web. *Proceedings of the Student Workshop of the 38th Annual Meeting of the Association for Computational Linguistics*, 29-36.

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

Karov, Y. and Edelman, S. (1996). Similarity-based Word Sense Disambiguation. *Computational Linguistics.* 24, 41-59.

Keller, F. and Lapata, M. (2003). Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics*, 3, 459- 484.

Keller, F., Lapata, M. and Ourioupina, O. (2002). Using the Web to Overcome Data Sparseness. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 230-237.

Kenney, C. and Boguraev, B. (1996). Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. *Proceedings of the 16th International Conference on Computational Linguistics (COLIN'96)*, Denmark, 113-118.

Korhonen, A. (2002). Subcategorization Acquisition. Technical Report, Cambridge University.

Khoja, S. (2001). APT: Arabic Part-of-Speech Tagger. *Proceedings of the Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, PA, USA.

Khoja, S., Garside, R., and Knowles, G. (2001). A Tagset for the Morpho-Syntactic Tagging of Arabic. *Proceedings of the Corpus Linguistics 2001 Conference*, Lancaster.

Kilgarriff, A. and Grefenstette, G. (2006). Web as Corpus. *Proceedings of Corpus Linguistics 2001*, Lancaster, 342 – 344.

Lappin, S. (2005). A Sequenced Model of Anaphora and Ellipsis Resolution. In B. Antonio, T. McEnery and R. Mitkov (Eds.). *Anaphora Processing: Linguistic, Cognitive and Computational Modeling* (pp. 3-17). Amsterdam; Philadelphia: J. Benjamins.

Lappin, S. and Less, H. (1994). An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics,* 20, 35-561.

Larkey, L., Ballesteros, L. and Connell, M. (2002). Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. *Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval*, 269-303.

Larkey, L. and Connell, M. (2002). Arabic Information Retrieval at UMass in TREC-10. *Proceedings of the 10th Text Retrieval Conference (TREC 2001)*, 562-570.

LDC, et al. (2004). *Arabic English Parallel News Part 1*. Linguistic Data Consortium, Philadelphia.

Liddy, E. (2001). Natural Language Processing. *Encyclopedia of Library and Information Science.* (pp. 2126 – 2136) 2nd Ed. New York: Marcel Decker, Inc.

Lu, X. (2005). Hybrid Methods for POS Guessing of Chinese Unknown Words. *Proceedings of the Student Research Workshop at the 43rd Annual Meeting of*

the Association for Computational Linguistics (ACL-05), (pp. 1-6). Ann Arbor, MI: The Association for Computational Linguistics.

Manaris, B. (1998). Natural Language Processing: A HUMAN–Computer Interaction Perspective. In Zelkowitz, M. (Ed.), *Advances in Computers*, Vol. 47 (pp. 1-66). New York: Academic Press.

Mani, N. (2004). The Role of Prosody in Parsing Ambiguous Sentences. *Proceedings of Speech Prosody 2004*, Nara, Japan.

Manning, C. and Schütze, H. (2002). *Foundations of Statistical Natural Language Processing*. London: The MIT Press.

McEnery, T. and Wilson, A. (2001). *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

McEnery, A. and Xiao, R. (2005). Parallel and Comparable Corpora: What are they up to? In James, G. and Anderman, G. (Eds.) *Incorporating Corpora: Translation and the Linguist* . Clevedon: UK.

Miller, G. (2005). English WordNet v.2.1. [Online]. Available: http://wordnet.princeton.edu

Mitkov, R. (1996). Anaphora and Machine Translation. *Machine Translation Review*, 4, 6-16.

Mitkov, R. (1997). Factors in Anaphora Resolution: They are not the only Things that Matter: A Case Study based on Two Different Approaches. *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, Madrid, Spain, 14-21.

Mitkov, R. (1998). Robust Pronoun Resolution with Limited Knowledge. *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, Montreal, Canada, 869-875.

Mitkov, R. (1999). Anaphora Resolution: The State of the Art. Technical Report based on *COLING'98* and *ACL'98* Tutorial on Anaphora Resolution, University of Wolverhampton.

Mitkov, R. (2001). Outstanding Issues in Anaphora Resolution. *Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing,* Mexico City, USA, 110-125.

Moore. R. (2004). On Log-Likelihood-Ratios and the Significance of Rare Events. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Myka, H. and Doucet, A. (2005). Data Mining Meets Collocations Discovery. In Arppe, A., Carlson, L., Lindén, K., Piitulainen, J., Suominen, M., Vainio, M., Westerlund, H. and Ylijyr, A. (Eds.). *Inquiries into Words, Constraints, and Contexts.* (pp. 194-203). Stanford, California: CSLI Publications.

Nelken, R. and Shieber, S. (2005). Arabic Diacritization Using Weighted Finite-State Transducers. *Proceedings of the 2005 Association for Computational Linguistics Workshop on Computational Approaches to Semitic Languages,* Ann Arbor, Michigan, 79-86.

Ng, V. and Cardie, C. (2002). Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution. *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002).*

Nunan, D. (1993). *Introducing Discourse Analysis.* England: Penguin Books Ltd.

Nwesri, A., Tahaghoghi, S. and Scholer, F. (2007). Capturing Out-Of-Vocabulary Words in Arabic Text. *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Prague, Czech Republic.

O'Hara, T., Bruce, R., Donner, F. and Wiebe, J. (2004). Class-based Collocations for Word-Sense Disambiguation. *Proceedings of Senseval 3 Workshop on Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London; New York: Routledge.

Olteanu, M. and Moldovan, D. (2005). PP-attachment Disambiguation using Large Context. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Canada.

Onaizan, Y. and Knight, K. (2002). Translating Named Entities Using Monolingual and Bilingual Resources. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, 400-408.

Pantel, P. and Lin, D. (2001). A Statistical Corpus-based Term Extractor. In Stroulia, E. and Matwin, S. (Eds.) *AI 2001, Lecture Notes in Artificial Intelligence*. (pp. 36-46). Springer-Verlag.

Pecina, P. and Schlesinger, P. (2006). Combining Association Measures for Collocation Extraction. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Sydney, Australia, 651–658.

Resnik, P. (1999). Mining the Web for Bilingual Texts. *Proceedings of the 37th Annual Meeting of the Association Computational Linguistics*, College Park, Maryland, USA, 527-534.

Richter, F. (2006). Introduction to Computational Linguistics [Online]. Available: www.sfs.uni-tuebingen.de Accessed: 16 July 2007.

Sadat, F. and Habash, N. (2006). Combination of Arabic Preprocessing Schemes for Statistical Machine Translation. *Proceedings of the 21st International*

Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, 1–8.

Sandra, W., Harvey, M. and Preston, K. (1996). Rule-Based Reference Resolution for Unrestricted Text Using Part-of-Speech Tagging and Noun Phrase Parsing. *Proceedings of the International Colloquium on Discourse Anaphora and Anaphora Resolution (DAARC)*, Lancaster, UK, 441-456

Sarkar, A. and Roeck, A. (2004). A Framework for Evaluating the Suitability of Non-English Corpora for Language Engineering. *Proceedings of Language Resources and Evaluation Conference (LREC)*, Lisbon, Spain.

Siekmann, J. (2007). Introduction to Artificial Intelligence [Online]. Available: http://www.ags.uni-sb.de Accessed: 18 July 2007.

Smadja, F., McKeown, K. and Hatzivassiloglou, V. (1996). Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, Vol. 22, 1-39.

Soon, W., Ng, T. and Lim, D. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27, 521-544

Stone, M. and Doran, C. (1996). Paying Heed to Collocations. *Proceedings of the 8th International Workshop on Natural Language Generation (INLG'96)*, Herstmonceux Castle, UK, 91-100.

Tanaka, T. (2002). Measuring the Similarity between Compound Nouns in Different Languages Using Non-Parallel Corpora. *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, 981-987.

Uryupina, O. (2006). Coreference Resolution With and Without Linguistic Knowledge. *Proceedings of Language Resources and Evaluation Conference (LREC 2006)*, Genoa, Italy, 893–898.

Villasenor-Pineda, L., Montes y Gomez, M., Perez-Coutino, M. and Vaufreydaz, D. (2003). A Corpus Balancing Method for Language Model Construction. *Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003)*, Mexico City, USA, 393-401.

Volk, M. (2001). Exploiting the WWW as a Corpus to Resolve PP Attachment Ambiguities. *Proceedings of Corpus Linguistics 2001*, Lancaster, UK.

Webster, J. and Ku, C. (1992). Tokenization as the Initial Phase in NLP. *Proceeding of International Conference of Computational Linguistics (COLIG'92),* Nantes, France, 1106-1110.

Werth, P. (1999). *Text Worlds: Representing Conceptual Space in Discourse*. New York: Longman.

Williams, S., Harvey, M. and Preston, K. (1996). Rule-Based Reference Resolution for Unrestricted Text Using Part-of-Speech Tagging and Noun Phrase Parsing. *Proceedings of the Internat8ional Colloquium on DAAR*, Lancaster, UK, 441-456.

Wright, W. (1981). *A Grammar of the Arabic Language.* Lebanon: Library of Lebanon.

Xu, J., Fraser, A. and Weischedel, R. (2002). Empirical Studies in Strategies for Arabic Retrieval. *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002),* Tampere, Finland, 269-274.

Glossary

Artificial Intelligence: It is defined as "a field of endeavor where computers and software programs are designed to mimic human reasoning and learning processes through the discovery of algorithms" (*Glossary*, 2008: 116). AI aims at developing machines, whose intelligence (i.e. their ability to process, understand and generate) mimics human intelligence.

Anaphora Resolution: Anaphora Resolution (AR) is the process of determining the antecedent of a given anaphor (Mitkov 2001).

Baseline Models: They are "the simplest implementation that one can think of" (Burch and Osborne 2003: 19). For instance, in Part-of-Speed tagging a baseline model can be a model that always assigns the most frequent tag given a particular word without attempting to do any contextual disambiguation.

Computational Linguistics: It is a subsection of Natural Language Processing (NLP) since it provides the models for the different linguistic phenomena that are used in NLP tasks. In other words, it is the working component of NLP systems**.**

Error Analysis: It is an analysis for the errors of the algorithm and their reasons. It indicates how the system may be improved in subsequent redesigns; being able to inspect the types of errors that are being made, and being able to generalize them into linguistic features is useful for the subsequent redesign of the statistical model used by the system (Burch and Osborne 2003).

Gold Standard: It is a measure of comparison that is considered ultimate or ideal.

Light Stemming: It is the "process of stripping off a small set of prefixes and/or suffixes, without trying to deal with infixes, or recognize patterns and find roots" (Larkey 2002: 276).

Natural Language Processing: It is "an area of research and application that explores how computers can be used to understand and manipulate written or spoken natural languages to perform such applications as Machine Translation (MT), Text Summarization (TS), Question Answering (QA), Information Retrieval (IR), Speech Recognition (SR) ... etc" (Chowdhury 2003: 51). Natural Language Processing is the subfield of Artificial Intelligence (AI) that deals with designing algorithms for computers to process, understand and generate language in the way humans do.

Part-of-Speech Tagging: It is the processing of labeling every word in the corpus with the correct part of speech.

Statistical Natural Language Processing: It is a Natural Language Processing approach, which relies on available data as its primary source of evidence to approximate generalized models of the linguistic phenomena under study based only on actual examples (Liddy 2001).

Appendixes

Appendix A

Distribution Tables of T-test Critical Values and X^2 Critical Values

(1) *T-Test* Critical Values (Manning and Schütze 2002)

P		0.05	0.025	0.01	0.005	0.001	0.0005
C		90%	95%	98%	99%	99.8%	99.9%
d.f.	1	6.314	12.71	31.82	63.66	318.3	636.6
	10	1.812	2.228	2.764	3.169	4.144	4.587
	20	1.725	2.086	2.528	2.845	3.552	3.850
(z)	∞	1.645	1.960	2.326	2.576	3.091	3.291

(2) X^2 Critical Values (Manning and Schütze 2002)

p	0.99	0.95	0.10	0.05	0.01	0.005	0.001
d.f. 1	0.00016	0.0039	2.71	3.84	6.63	7.88	10.83
2	0.020	0.10	4.60	5.99	9.21	10.60	13.82
3	0.115	0.35	6.25	7.81	11.34	12.84	16.27
4	0.297	0.71	7.78	9.49	13.28	14.86	18.47
100	70.06	77.93	118.5	124.3	135.8	140.2	149.4

Appendix B

Frequent Words used to Estimate the Web Size

أثناء	أيا	إنك	الذي	اياها	بعضها	بيني	خلال	عليه	ف
أحيانا	أية	إنما	الذين	ب	بعضهم	تارة	خلاله	عليها	فأن
ألا	أين	إننا	اللائي	بأن	بغير	تبا	خلالها	عليهم	فإن
أما	أينما	إنني	اللاتي	بأنك	بـ	تجاه	دائما	عما	فإننا
أمس	أيهما	إنه	اللتان	بأننا	بـ	تجاهها	دعك	عن	فإنني
أن	إبان	إنها	اللواتي	بأنني	بك	تلك	دون	عند	فإنه
أنا	إثر	إنهم	اللي	بأنه	بكل	تماما	دونما	عندئذ	فإنها
أنت	إذ	إنهما	إلى	بأنها	بل	ثم	ذلك	عندك	فإنهم
أنتم	إذا	إني	إلي	بأنهم	بلا	ثمة	ربما	عندكم	فإني
أنك	إزاء	إياه	إليه	بأنهما	بما	جراء	رغم	عندما	فتكون
أننا	إلا	ا	اليوم	بأي	بماذا	حاليا	سواء	عنده	فتكونوا
أنني	إلى	إذ	أما	بأن	بنا	حتى	سوف	عندها	فجأة
أنه	إلي	إذا	أمس	بتلك	به	حتي	ضد	عندي	فـي
أنها	إلينا	إزاء	أن	بدون	بها	حسب	طالما	عنك	فقد
أنهم	إليه	ال	انت	برغم	بهذا	حول	طيلة	عنه	فقط
أنهما	إليها	الآن	انك	بعد	بهذه	حيال	على	عنها	فكان
أني	إليهم	التى	إنما	بعدما	بي	حيث	علي	عنهم	فكانت
أو	إما	التي	إنه	بعض	بين	حين	عليك	غدا	فلا
أولئك	إن	الدي	أنها	بعضكم	بينما	حينا	عليكم	غير	فلو
أي	إنا	الذى	أو	بعضه	بينهما	حينما	علينا	ـإن	فهل

فهو	كل	لابد	له	مثلها	مهما	و	وإلى	وكذلك	ومنها
فوق	كلاهما	لاسيما	لهؤلاء	مذ	نا	وأن	وإلي	وكل	وهذا
فى	كلما	لانها	لها	مرة	ناهيك	وأننا	وإن	وكيف	وهذه
في	كلهم	لدى	لهم	مع	نحن	وأنني	وإنما	ولا	وهكذا
فيكن	كلهن	لدي	لهن	معا	ه	وأنه	وبأن	ولاسيما	وهل
فيكون	كم	لديها	لو	معك	هؤلاء	وأنها	وبل	ولقد	وهنا
فيكونوا	كما	لعل	لولا	معكم	ها	وأنهما	وبين	ولكن	وهو
فيما	كى	لـ	لي	معه	هذا	وإذا	وتكون	ولم	وهي
فيه	كي	لقد	ليته	معها	هذه	وإلا	وتكونوا	ولماذا	ويكن
فيها	كيف	لك	ليس	معهم	هذين	وإن	وحتى	ولنا	ويكون
فيهم	ل	لكل	م	مما	هكذا	وإنا	وحتي	وله	ويكونوا
قبل	لأجل	لكن	ما	ممن	هل	وإنما	وخلال	ولها	ويل
قد	لأن	لكنه	مابين	من	هم	وإنني	وذلك	ولهذا	ويلي
ك	لأنني	لكي	مادمت	منا	هما	وإنه	وسوف	ولهم	يا
كأن	لأنه	لم	ماذا	منذ	هن	وإنها	وعلينا	ولو	يكن
كأنه	لأنها	لما	مامن	منه	هنا	وإياها	وغالبا	وليس	يكون
كأنها	لأنهما	لماذا	ماهو	منها	هناك	وإلا	وفي	وماذا	يوم
كان	لأول	لمن	متى	منهم	هنالك	والتي	وقد	ومع	يومـــــا
كانت	لأي	لن	مثل	منهما	هو	والذي	وكان	ومعهم	
كذلك	لا	لنا	مثلما	منهن	هي	والذين	وكانت	ومن	

Appendix C

Perl Codes Used Throughout the Thesis

```
##############################################################################################
## Creator: Khaled Elghamry
## Date: 20 December 2006
## Usage: Corpus-based Tokenizer
## Platform: Windows XP
##############################################################################################

#### Step 1: Hashing the entire corpus
@files = glob('raw_text.txt');                    ### Raw Arabic Text
for $f(@files){
open(F, $f);
while(<F>){
require 'cat.lib'; &clean; &d;                    ### cleans numbers and punctuation markers
s/ ال/ DET<ال> /g;
@w = split;
for $x(@w){
++$all{$x}}}}

#### Step 2: Tokenization
@files = glob('raw_test.txt');                    ### Raw Arabic Text
for $f(@files){
open(F, $f);
while(<F>){
require 'cat.lib'; &clean; &d;                    ### cleans numbers and punctuation markers
s/ ال/ DET<ال> /g;
s/ لل/ PREP<ل> DET<ال> /g;
s/ به / PREP<ب> PRO<ه> /g;
s/ بها / PREP<ب> PRO<ها> /g;
s/ بنا / PREP<ب> PRO<نا> /g;
s/ له / PREP<ل> PRO<ه> /g;
s/ لها / PREP<ل> PRO<ها> /g;
s/ لنا / PREP<ل> PRO<نا> /g;

@w1 = split;
for $y(@w1){
$conj = substr($y,0,1);
$det = substr($y,1,2);
$stem = substr($y,3);
if(substr($y,0,3) eq 'وال' or substr($y,0,3) eq 'فـال' and exists $all{$stem} and
length($stem) >1){
s/ $y / CONJ<$conj>DET<$det> $stem /g;}}

@w2 = split;
for $y(@w2){
$prep = substr($y,0,1);
$det = substr($y,1,2);
$stem = substr($y,3);
if(substr($y,0,3) eq 'بـال' or substr($y,0,3) eq 'كـال' and exists $all{$stem} and
length($stem) >1){
```

```
s/ $y / PREP<$prep>DET<$det> $stem /g;}}

@w3 = split;
for $y(@w3){
$prep = substr($y,0,1);
$stem = substr($y,1);
if($prep =~ /ك|ب|ل/ and exists $all{$stem} and length($stem) >1){
s/ $y / PREP<$prep> $stem /g;}}

@w4 = split;
for $y(@w4){
$conj = substr($y,0,1);
$stem = substr($y,1);
if($conj =~ /ف|و/ and exists $all{$stem} and length($stem) >1){
s/ $y / CONJ<$conj> $stem /g;}}

@w5 = split;
for $y(@w5){
$conj = substr($y,0,1);
$prep = substr($y,1,1);
$stem = substr($y,2);
if($conj =~ /ف|ل|ك|ب|و/ and exists $all{$stem} and length($stem) >1){
s/ $y / CONJ<$conj>PREP<$prep> $stem /g;}}

@w6 = split;
for $y(@w6){
$pro = substr($y,-2);
$stem = substr($y,0,-2);
if($pro =~ /ها|هم|هن|نا/ and exists $all{$stem} and length($stem) >1){
s/ $y / $stem PRO<$pro> /g;}}

@w7 = split;
for $y(@w7){
$pro = substr($y,-1);
$stem = substr($y,0,-1);
if($pro =~ /ك|ه/ and exists $all{$stem} and length($stem) >1){
s/ $y / $stem PRO<$pro> /g;}}

@w8 = split;
for $y(@w8){
$pro = substr($y,-1);
$stem = substr($y,0,-1);
if($pro =~ /ا/ and exists $all{$stem} and length($stem) >1){
s/ $y / $stem ACC<$pro> /g;}}
&cleaning;
s/ /\n/g;
print G }}

#### Cleans some of the common mistakes
sub cleaning
{s/DET<ال> CONJ<و> /DET<ال> و/g;
s/ DET<ال> تي / التي /g;
s/ اذ ACC<ا> / اذا /g;
```

s/ DET<ال> ذي / الذي /g;}

```
##############################################################################################
## Creator: Rania Al-Sabbagh
## Date: 20 December 2006
## Usage: Corpus-based and Dictionary-Based Tokenizer
## Platform: Windows XP
##############################################################################################

#### Step 1: Hashing raw Arabic text
@files = glob('raw_text.txt');          ### Raw Arabic Text
for $f(@files){
open(F, $f);
while(<F>){
require 'cat.lib'; &clean; &d;          ### cleans numbers and punctuation markers
@w = split;
for $x(@w){
++$all{$x}}}};

#### Step 2: Hashing Arabic text analyzed according to Buckwalter's AraMorph (2002)
@file = glob('bama_text.txt');          ### Arabic text analyzed by AraMorph
for $f(@file){
open(S, $f);
while($line = <S>){
@w = split(' ', $line);

for $r(0..@w){
if(@w[$r] =~ /DET/ and @w[$r] !~ /CONJ|PREP/){
$e = substr(@w[0],2);
++$al{$e}}}                             ### Definite words after removing the definite article

for $r(0..@w){
if(@w[$r] =~ /DET/ and @w[$r] =~ /CONJ/ and @w[$r] !~ /PREP/){
$k = substr(@w[0],3);
++$conj_al{$k}}}                        ### Definite words procliticzed to a conjunction after
                                        removing the definite article and the conjunction

for $r(0..@w){
if(@w[$r] =~ /DET/ and @w[$r] =~ /PREP/ and @w[$r] !~ /CONJ/){
$w = substr(@w[0],3);
++$prep_al{$w}}}                        ### Definite words procliticzed to a preposition after
                                        removing the definite article and the preposition

for $r(0..@w){
if(@w[$r] =~ /PREP/ and @w[$r] !~ /DET|CONJ/){
$x = substr(@w[0],1);
++$prep_only{$x}}}                      ### Words procliticzed to a preposition after removing
                                        the definite article and the preposition

for $r(0..@w){
if(@w[$r] =~ /CONJ/ and @w[$r] !~ /DET|PREP/){
$q = substr(@w[0],1);
++$conj_only{$q}}}                      ### Words procliticzed to a conjunction after removing
                                        the definite article and the conjunction

for $r(0..@w){
if(@w[$r] =~ /PREP/ and @w[$r] =~ /CONJ/ and @w[$r] !~ /DET/){
$c = substr(@w[0],2);
```

```
++$conj_prep{$c}}}                    ### Words encliticzed to a conjunction and a preposition after removing both

for $r(0..@w){
if(@w[$r] =~ /\[CONJ_PREP\]/){
$con = substr(@w[0],0,1);
$pre = substr(@w[0],1);
++$conj_prep2{$con};
++$conj_prep3{$pre}}}                 ### Words that consist of a conjunction and a preposition

for $r(0..@w){
if(@w[$r] =~ /DET/ and @w[$r] =~ /CONJ/ and @w[$r] =~ /PREP/){
++$every{@w[0]}}}                     ### Definite Words procliticzed to a conjunction and a preposition removing all

for $r(0..@w){
if(@w[$r] =~ /3FS|3FP|3MP/){
$n = substr(@w[0],0,-2);
if(substr($n,-1) eq 'ت' and not exists $all{$n}){
$n =~ s/ة/ت/g;}
if(substr($n,-1) eq 'ئ' and not exists $all{$n}){
$n =~ s/ئ/ء/g;}
++$pronoun{$n}}};                     ### Words encliticzed to the pronouns ها, هن, هم, نا

for $r(0..@w){
if(@w[$r] =~ /3MS|2MS/){
$o = substr(@w[0],0,-1);
if(substr($o,-1) eq 'ت' and not exists $all{$o}){
$o =~ s/ة/ت/g;}
if(substr($o,-1) eq 'ئ and not exists $all{$o}){
$o =~ s/ئ/ء/g;}
++$pronoun2{"$o"}}}                   ### Words encliticzed to the pronouns ه, ك

for $r(0..@w){
if(@w[$r] =~ /3D/){
$l = substr(@w[0],0,-3);
if(substr($l,-1) eq 'ت' and not exists $all{$l}){
$l =~ s/ة/ت/g;}
if(substr($l,-1) eq 'ئ' and not exists $all{$l}){
$l =~ s/ئ/ء/g;}
++$pronoun3{$l}}}                     ### Words encliticzed to the pronoun هما

@files = glob('raw_text.txt');        ### Raw text to be tokenized
for $f(@files){
open(F, $f);
while(<F>){
require 'cat.lib'; &clean; &d;        ### cleans numbers and punctuation markers
s/ به / PREP<ب> PRO<ه> /g;
s/ بها / PREP<ب> PRO<ها> /g;
s/ بنا / PREP<ب> PRO<نا> /g;
s/ له / PREP<ل> PRO<ه> /g;
s/ لها / PREP<ل> PRO<ها> /g;
s/ لنا / PREP<ل> PRO<نا> /g;
```

```
@w1 = split;
for $y(@w1){
$conj = substr($y,0,1);
$det = substr($y,1,2);
$stem = substr($y,3);
if(substr($y,0,3) eq 'وال' or substr($y,0,3) eq 'فال' and exists $all{$stem} and exists $conj_al{$stem} and length($stem) >1){
s/ $y / CONJ<$conj>_DET<$det>_$stem /g;}}

@w2 = split;
for $y(@w2){
$prep = substr($y,0,1);
$det = substr($y,1,2);
$stem = substr($y,3);
if(substr($y,0,3) eq 'بال' or substr($y,0,3) eq 'كال' and exists $all{$stem} and  exists $prep_al{$stem} and length($stem) >1){
s/ $y / PREP<$prep>_DET<$det>_$stem /g;}}

@w3 = split;
for $y(@w3){
$prep = substr($y,0,1);
$det = substr($y,1,1);
$stem = substr($y,2);
if(substr($y,0,2) eq 'لل' or substr($y,0,2) eq 'للا' and exists $all{$stem} and  exists $prep_al{$stem} and length($stem) >1){
s/ $y / PREP<$prep>_DET<$det>_$stem /g;}}          ### new addition

@w4 = split;
for $y(@w4){
$prep = substr($y,0,1);
$stem = substr($y,1);
if($prep =~ /ك|ب|ل/ and exists $all{$stem} and exists $prep_only{$stem} and length($stem) >1){
s/ $y / PREP<$prep>_$stem /g;}};

@w5 = split;
for $y(@w5){
$conj = substr($y,0,1);
$stem = substr($y,1);
if($conj =~ /ف|و/ and exists $all{$stem} and  exists $conj_only{$stem} and length($stem) >1){
s/ $y / CONJ<$conj>_$stem /g;}}

@w6 = split;
for $y(@w6){
$conj = substr($y,0,1);
$prep = substr($y,1,1);
$stem = substr($y,2);
if($conj =~ /ك|ف|ل|ب|و/ and exists $all{$stem} and exists $conj_prep{$stem} and length($stem) >1){
s/ $y / CONJ<$conj>_PREP<$prep>_$stem /g;}}

@w7 = split;
```

```
for $y(@w7){
$conj = substr($y,0,1);
$prep = substr($y,1);
if(exists $all{$conj} and exists $all{$prep} and exists $conj_prep2{$conj} and exists $conj_prep3{$prep} and length($prep) >1){
s/ $y / CONJ<$conj>_PREP<$prep> /g;}}          ### new addition

@w8 = split;
for $y(@w8){
$pro = substr($y,-2);
$stem = substr($y,0,-2);
if(substr($stem,-1) eq 'ت' and not exists $all{$stem}){
$stem =~ s/ت/ة/g; }
if(substr($stem,-1) eq 'ئ' and not exists $all{$stem}){
$stem =~ s/ئ/ء/g;}
if($pro =~ /نا|ها|هم|هن/ and exists $all{$stem} and exists $pronoun{$stem} and length($stem) >1){
s/ $y / <$stem>_PRO<$pro> /g;}}

@w9 = split;
for $y(@w9){
$pro = substr($y,-1);
$stem = substr($y,0,-1);
if(substr($stem,-1) eq 'ت' and not exists $all{$stem}){
$stem =~ s/ت/ة/g; }
if(substr($stem,-1) eq 'Æ' and not exists $all{$stem}){
$stem =~ s/ئ/ء/g;}
if($pro =~ /ه/ and exists $all{$stem} and exists $pronoun2{$stem} and length($stem) >1){
s/ $y / <$stem>_PRO<$pro> /g;}}

@w10 = split;
for $y(@w10){
$pro = substr($y,-3);
$stem = substr($y,0,-3);
if(substr($stem,-1) eq 'ت' and not exists $all{$stem}){
$stem =~ s/ت/ة/g;}
if(substr($stem,-1) eq 'ئ' and not exists $all{$stem}){
$stem =~ s/ئ/ء/g;}
if($pro =~ /هما/ and exists $all{$stem} and exists $pronoun3{$stem} and length($stem) >1){
s/ $y / <$stem>_PRO<$pro> /g;}}          ### new addition

@w11 = split;
for $y(@w11){
if(substr($y,0,2) eq 'ال'){
$det = substr($y,0,2);
$stem = substr($y,2);
if(exists $all{$stem} and exists $all{$stem} and length($stem) >1){
s/ $y / DET<$det>_<$stem> /g;}}};
&cleaning;
s/ /\n/g;
```

```
print G }}

sub cleaning
{
s/DET<ال> ي / الي /g;                    ### new additions
s/PREP<ك> كلمة /لمة /g;
s/DET<ال> تي / التي /g;
s/DET<ال> ذي / الذي /g;
s/DET<ال> <له> / الله/g;
s/DET<ال> ذين / الذين /g;
s/PREP<ب> ين / بين /g;
s/PREP<ك> ان / كان /g;
s/PREP<ب> عد /بعد /g;}
```

```
##############################################################################################
## Creator: Rania Al-Sabbagh
## Date: 25 May 2007
## Usage: To calculate the conditional probability of the bigrams whose counts are taken from the Web
## Platform: Windows XP
##############################################################################################

$Web_size = 4500000000;                 ### the estimated Web size

#### Step 1: Hashing bigrams and unigrams and getting their Web counts
open (F, 'Web_counts.txt');             ### Web's counts for each bigram & unigram
while($line = <F>){
@w = split(' ', $line);
$l = @w;
if($l > 2){
++$big{"@w[0] @w[1] @w[2]"}}    ### a hash of the bigrams
if($l == 2){
++$uni{"@w[0] @w[1]"} }}                ### a hash of the unigrams

#### Step 2: Calculating the probability of each bigram and unigrams based on their Web counts
#### CP is calculating as follows: P (antecedent|carrier) = P (bigram)/ P(carrier)
for $x(sort keys %big){
($a, $b, $c) = split(' ', $x);          ### $a=  bigram's count $b= antecedent  $c= carrier

for $t(sort keys %uni){
($d, $e) = split(' ', $t);              ### $d=  unigram's count $e= the unigram

if($e eq $c){
$a =~ s/\,//g;
$d =~ s/\,//g;

$p12 = $a / $Web;                               ### probability of each bigram
$p1 = $d / $Web;                                ### probability of each unigram

$con = $p12 / ($p1+0.00000000000000000000000000000001);
print  "$b $c \t"; printf  "%2.10f" ,"$con"; print  "\n";}}}
```

```
##############################################################################################
## Creator: Rania Al-Sabbagh
## Date: 27 May 2007
## Usage: The AR algorithm
## Platform: Windows XP
##############################################################################################

#### Step1: The conditional probability of the bigrams (i.e. of the carrier of the pronoun and each candidate antecedent)
@files = glob('con_prop.txt');
for $f(@files){
open(F, $f);
while($line = <F>){
push(@out, $line)}}

#### Step 2: Hashing unigrams, bigrams and bands
for $l(@out){
@h = split(' ', $l);
$l1 = @h;
if($l1 = = 2){                          ### unigrams
$j = @h[1];
$mod{$j} = @h[0]}

if($l1 = = 3){
$j2 = join(' ', @h[1], @h[2]);          ### bigrams
$mod2{$j2} = @h[0]}

if($l1 > 3){
$j3 = join(' ', @h[1..$l1]);            ### bands
$mod3{$j3} = @h[0]}}

#### Step 3: Hashing the conditional probability model
for $x(keys %mod2){
($w1, $w2) = split(' ', $x);
$score = $mod2{$x};
$bag = join(' ', $x, $score);
push(@scores, $bag)}

for $s(@scores){
@hh = split(' ', $s);
$m = join(' ', @hh[0], @hh[1]);
$mod{$m} = @hh[2]}

#### Step 4: Dividing the Arabic sentences into bands and filtering them according to their cumulative counts
open(G, 'arabic_sentences.txt');
while($line2 = <G>){
@t = split(' ',$line2);

#### Dividing the 20-word window size into two equal chunks A & B
for $j(0..9){
push(@bag1, @t[$j]);
$bi1 = join(' ', @t[$j], @t[19]);
$score1 += $mod{$bi1}}
print "@bag1 $score1\n";
```

```
for $j(10..19){
push(@bag2, @t[$j]);
$bi2 = join(' ', @t[$j], @t[19]);
$score2 += $mod{$bi2}}
print "@bag2 $score2\n";
```

If score A is higher than score B, the A chunk is further subdivided into two equal chunks C & D

```
if($score1 > $score2){
for $j(0..4){
push(@bag3, @t[$j]);
$bi3 = join(' ', @t[$j], @t[19]);
$score3 += $mod{$bi3}}
print "@bag3 $score3\n";

for $j(5..9){
push(@bag4, @t[$j]);
$bi4 = join(' ', @t[$j], @t[19]);
$score4 += $mod{$bi4}}
```

If score C is higher than score D, the C chunk is divided into bigrams and the bigram with the highest count is selected

```
if($score3 > $score4){
for $j(0..1){
push(@bag5, @t[$j]);
$bi5 = join(' ', @t[$j], @t[19]);
$score5 += $mod{$bi5}}
print "@bag5 $score5\n";

for $j(1..2){
push(@bag6, @t[$j]);
$bi6 = join(' ', @t[$j], @t[19]);
$score6 += $mod{$bi6}}
print "@bag6 $score6\n";

for $j(2..3){
push(@bag7, @t[$j]);
$bi7 = join(' ', @t[$j], @t[19]);
$score7 += $mod{$bi7}}
print "@bag7 $score7\n";

for $j(3..4){
push(@bag8, @t[$j]);
$bi8 = join(' ', @t[$j], @t[19]);
$score8 += $mod{$bi8}}
print "@bag8 $score8\n"}
```

If score D is higher than score C, the D chunk is divided into bigrams and the bigram with the highest count is selected

```
if($score4 > $score3){
for $j(5..9){
push(@bag9, @t[$j]);
$bi9 = join(' ', @t[$j], @t[19]);
$score9 += $mod{$bi9}}
print "@bag9 $score9\n";
```

```
for $j(5..6){
push(@bag10, @t[$j]);
$bi10 = join(' ', @t[$j], @t[19]);
$score10 += $mod{$bi10}}
print "@bag10 $score10\n";

for $j(6..7){
push(@bag11, @t[$j]);
$bi11 = join(' ', @t[$j], @t[19]);
$score11 += $mod{$bi11}}
print "@bag11 $score11\n";

for $j(7..8){
push(@bag12, @t[$j]);
$bi12 = join(' ', @t[$j], @t[19]);
$score12 += $mod{$bi12}}
print "@bag12 $score12\n";

for $j(8..9){
push(@bag13, @t[$j]);
$bi13 = join(' ', @t[$j], @t[19]);
$score13 += $mod{$bi13}}
print "@bag13 $score13\n"}}
```

If score B is higher than score A, the B chunk is further subdivided into two equal chunks E & F

```
if($score2 > $score1){
for $j(10..14){
push(@bag14, @t[$j]);
$bi14 = join(' ', @t[$j], @t[19]);
$score14 += $mod{$bi14}}
print "@bag14 $score14\n";

for $j(15..19){
push(@bag15, @t[$j]);
$bi15 = join(' ', @t[$j], @t[19]);
$score15 += $mod{$bi15}}
print "@bag15 $score15\n";
```

If score E is higher than score F, the E chunk is divided into bigrams and the bigram with the highest count is selected

```
if($score14 > $score15){
for $j(10..11){
push(@bag16, @t[$j]);
$bi16 = join(' ', @t[$j], @t[19]);
$score16 += $mod{$bi16}}
print "@bag16 $score16\n";

for $j(11..12){
push(@bag17, @t[$j]);
$bi17 = join(' ', @t[$j], @t[19]);
$score17 += $mod{$bi17}}
print "@bag17 $score17\n";

for $j(12..13){
push(@bag18, @t[$j]);
```

```
$bi18 = join(' ', @t[$j], @t[19]);
$score18 += $mod{$bi18}}
print "@bag18 $score18\n";

for $j(13..14){
push(@bag19, @t[$j]);
$bi19 = join(' ', @t[$j], @t[19]);
$score19 += $mod{$bi19}}
print "@bag19 $score19\n";}

#### If score F is higher than score E, the F chunk is divided into bigrams and the bigram with the highest count is selected
if($score15 > $score14){
for $j(15..16){
push(@bag20, @t[$j]);
$bi20 = join(' ', @t[$j], @t[19]);
$score20 += $mod{$bi20}}
print "@bag20 $score20\n";

for $j(16..17){
push(@bag21, @t[$j]);
$bi21 = join(' ', @t[$j], @t[19]);
$score21 += $mod{$bi21}}
print "@bag21 $score21\n";

for $j(17..18){
push(@bag22, @t[$j]);
$bi22 = join(' ', @t[$j], @t[19]);
$score22 += $mod{$bi22}}
print "@bag22 $score22\n";

for $j(18..19){
push(@bag23, @t[$j]);
$bi23 = join(' ', @t[$j], @t[19]);
$score23 += $mod{$bi23}}
print "@bag23 $score23\n";}}

@t = (); @bag1 = ();@bag2 = ();@bag3 = ();@bag4 = ();@bag5 = ();@bag6 = ();@bag7 = ();@bag8 = ();@bag9 = ();@bag10 = ();@bag12 = ();@bag13 = ();@bag14 = ();@bag15 = ();@bag16 = ();@bag17 = ();@bag18 = ();@bag19 = ();@bag20 = ();@bag21 = ();@bag22 = ();@bag23 = ();

$score1 = 0;$score2 = 0;$score3 = 0;$score4 = 0;$score5 = 0;$score6 = 0;$score7 = 0;$score8 = 0;$score9 = 0;$score10 = 0;$score11 = 0;$score12 = 0;$score13 = 0;$score14 = 0;$score15 = 0;$score16 = 0;$score17 = 0;$score18 = 0;$score19 = 0;$score20 = 0;$score21 = 0;$score22 = 0;$score23 = 0;

print "$line2\n";}
```

Appendix D

Some Arabic Corpora frequently used in ANLP

Corpus Name	Source	Description	Status
Arabic Broadcast News Transcripts	Linguistic Data Consortium (LDC)	This data set consists of eight text files containing transcripts for Voice of America satellite radio news broadcasts in Arabic. The broadcasts were recorded by the Linguistic Data Consortium at transmission time between June 2000 and January 2001.	Non-Free
Egyptian Colloquial Arabic Lexicon	Linguistic Data Consortium (LDC)	This lexicon represents the first electronic pronunciation dictionary of Egyptian Colloquial Arabic (ECA), the spoken variety of Arabic found in Egypt. The dialect of ECA that this dictionary represents is Cairene Arabic	Non-Free
Arabic Broadcast News Speech	Linguistic Data Consortium (LDC)	This data set consists of eight audio files recorded by the Linguistic Data Consortium from Voice of America satellite radio news broadcasts in Arabic. The recordings were made at time of transmission between June 2000 and January 2001. This work was sponsored in part by National Science Foundation Grant No. IIS-9982201.	Non-Free
GALE Phase 1 Arabic Broadcast News Parallel Text - Part 1	Linguistic Data Consortium (LDC)	This release is Part 1 of the three-part GALE Phase 1 Arabic Broadcast News Parallel Text, which, along with other corpora, was used as training data in year 1 (Phase 1) of the DARPA-funded GALE program. This corpus contains transcripts and English translations of 17 hours of Arabic broadcast news programming selected from a variety of sources.	Non-Free
Arabic English Parallel News Part 1	Linguistic Data Consortium (LDC)	This corpus contains Arabic news stories and their English translations LDC collected via Ummah Press Service from January 2001 to September 2004. It totals	Non-Free

		8,439 story pairs, 68,685 sentence pairs, 2M Arabic words and 2.5M English words. The corpus is aligned at sentence level. All data files are SGML documents.	
Arabic Newswire Part 1	Linguistic Data Consortium (LDC)	This publication contains the Arabic Newswire A Corpus, Linguistic Data Consortium (LDC) catalog number LDC2001T55 and ISBN 1-58563-190-6. The Arabic Newswire Corpus is composed of articles from the Agence France Presse (AFP) Arabic Newswire. The source material was tagged using TIPSTER-style SGML and was transcoded to Unicode (UTF-8). The corpus includes articles from May 13, 1994 to December 20, 2000.	
Arabic Treebank: Part 1 - 10K-word English Translation		Arabic Treebank: Part 1 - 10K-word English Translation was produced by Linguistic Data Consortium (LDC) catalog number LDC2003T07 and ISBN 1-58563-262-7. The purpose of this corpus of 10K Arabic words translated into English is to support the development of data-driven approaches to natural language processing, machine translation, human language technologies, cross-lingual information retrieval, and other forms of linguistic research on Modern Standard Arabic in general.	
Arabic Treebank: Part 1 v 3.0 (POS with full vocalization + syntactic analysis)		The project targets the description of a written Modern Standard Arabic corpus from the Agence France Presse (AFP) newswire archives for July-November 2000 (files dated 20000/7/15 to 2000/11/15). This corpus includes 734 stories representing 145,386 words (166,068 tokens after clitic segmentation in the Treebank; the number of Arabic tokens is 123,796). For this work, annotators must be native speakers of Arabic, and they must understand enough linguistics to check morphosyntactic analysis and build syntactic	