English to Arabic Machine Translation Using a Phrase-based Approach

BY

Mohammad Ismail Hasan Amro

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER SCIENCE

January 2012

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **Mohammad Ismail Hasan Amro** under the direction of his thesis advisors and approved by his thesis committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE**

Thesis Committee

Dr. Wasfi G. Al-Khatib
(Chairman)

M. Elsh

Dr. Moustafa Elshafei
(Member)

Dr. Husni Al-Muhtaseb
(Member)

عادل

Dr. Adel Ahmed
Department Chairman

Dr. Salam A. Zummo
Dean of Graduate Studies

23/1/12

Date

Dedicated to

My Parents

Whose Prayers and Perseverance led to this accomplishment

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and the Most Merciful. All praises and glory be to Allah (SWT) for blessing me with opportunities abound and showering upon me his mercy and guidance throughout my life. I pray that He continues the same the rest of my life. And may His peace and blessings of Allah be upon Prophet Muhammad, who has been a guidance and inspiration to our lives.

I would like to thank my supervisor, Dr. Wasfi G. Al-Khatib for his guidance throughout this thesis. He was always there when I needed him, despite his tight schedule. I am extremely grateful to him for his prompt replies and his numerous proofreads. I am also very grateful to my thesis committee members, Dr. Husni Al-Muhtaseb and Prof. Moustafa Elshafei, for their care, cooperation and constructive advice.

Special thanks go to my colleagues and friends for their encouragements and various help that they provided throughout my graduate studies at KFUPM. I would like to give my special thanks to my parents, my brother and my sister for their support, patience and love. Without their encouragement, motivation and understanding, it would have been impossible for me to complete this work.

Table Of Content

LIST OF TABLES

LIST OF FIGURES

THESIS ABSTRACT

Name: Mohammad Ismail Hasan Amro

Title: English to Arabic Machine Translation Using a Phrase-based Approach

Major Field: COMPUTER SCIENCE

Date of Degree: January 2012

Statistical machine translation (SMT) treats the translation of natural language as a machine learning problem. By examining many samples of human-produced translations, SMT algorithms automatically learn how to translate. In this thesis, we discuss the automatic machine translation from English to Arabic using a statistical phrase-based approach employing a parallel Arabic-English corpus that was developed manually by more than one translator. Statistical machine translation (SMT) consists of two phases: The training phase and the decoding phase. In the training phase, the statistical language model and the translation model are built. In the decoding phase, the best possible translation is chosen depending on a comprehensive search process. We built a sizable parallel corpus spanning various categories of topics from the Meedan website, and later compared the results of Meedan with that of the other two corpora: LDC and UN. The performance was compared based on the Bilingual Evaluation Understudy (BLEU). Our experimentation shows that, overall, the Meedan corpus outperformed the other two corpora in most categories. We, also, compared the performance of the Moses decoder and the Pharaoh decoder. We conclude that although the response time for the pharaoh

decoder is better than that of the Moses decoder, the quality of the translation of the Moses decoder exceeds that of the Pharaoh decoder.

ملخص الرسالة

الاســــــــــم : محمد إسماعيل حسن عمرو

عنوان الرسالة:الترجمة الآلية من الإنجليزية إلى العربية باستخدام الطريقة الإحصائية اعتمادا على المقاطع اللغوية المستخلصة من مكانز ثنائية اللغة

التخصـــص :علم حاسوب

تاريخ التخرج: يناير 2012

الخلاصة. تعد ترجمة اللغة الطبيعية بواسطة الطرق الإحصائية مسألة من مسائل التعلم الآلي ، حيث تتعلم خوارزميات الطريقة الإحصائية كيفية الترجمة تلقائيا بالاعتماد على العبارات اللغوية المترجمة بطريقة يدوية من عدة مترجمين. نناقش في هذه الأطروحة استخدام الطريقة الإحصائية المعتمدة على العبارات اللغوية في عملية الترجمة من اللغة الإنجليزية إلى اللغة العربية باستخدام مكنز ثنائي اللغة. تتكون هذه الطريقة الإحصائية في عملية الترجمة من مرحلتين رئيسيتين هما مرحلة التدريب ومرحلة النقل. يتم في المرحلة الأولى بناء النموذج الإحصائي اللغوي ونموذج الترجمة، ويتم في المرحلة الثانية اختيار أجود ترجمة ممكنة اعتمادا على عملية البحث الشامل. قمنا من خلال هذه الأطروحة ببناء مكنز لغوي يغطي مواضيع متنوعة بالاعتماد على البيانات اللغوية المستخلصة من موقع ميدان. كما قمنا بمقارنة نتائج الترجمة المعتمده على مكنز ميدان مع نتائج مكانز لغوية اخرى مثل مكنز الامم المتحدة ومكنز اتحاد البيانات اللغوية في جامعة بنسلفانيا LDC. حيث تمت هذه المقارنة باستخدام أداء BLEU . وخلال التجارب تبين لنا ان مكنز ميدان عموما

يتفوق على المكانز الاخرى . أيضا قمنا بالمقارنة بين Moses decoder و Pharaoh decoder
.واستنتجنا أنه بالرغم من أن وقت استجابة Pharaoh decoder هو أفضل من Moses decoder
، إلا أن جودة الترجمة عند استخدام Moses decoder أفضل من Pharaoh decoder.

blank

CHAPTER 1

Introduction

1.1. Overview

English is widely used in media, higher education and publishing. Translation of English content into Arabic greatly helps Arabic speakers to benefit from the immense English literature and web content available. The size of English Web content, in particular, is far larger than that of the available Arabic content. Therefore, developing effective English to Arabic machine translation systems is instrumental in promoting knowledge transfer and dissemination among Arab speakers. Since the September 11, 2001 events, the Arabic language received intensified attention from researchers in natural language processing, document analysis, optical character recognition, etc. [1]. Machine translation (MT) is defined as the process that utilizes computers to translate text from one natural language to another. Translation, in its full generality, is a difficult, attractive, and powerfully human effort, as rich as any other area of human creativity [2]. Difficulty for the translator from one language to another arises from the source language. It may have more words than the target translation language. For example, English has more words than Chinese. These differences are caused, in part, by the structural differences between the two languages [2]. Therefore, a lot of challenges need to be addressed when attempting machine translation. Machine translation approaches can be classified into classical MT approaches and statistical machine translation (SMT) approaches.

Classical MT approaches can be divided into three categories: the direct translation approach, the transfer approach and the Interlingua approach [2]. In direct translation, the text of the source language is scanned and translated word by word using a large bilingual dictionary. This dictionary contains simple morphological rules associated with each word that will aid in the translation process. After the words are translated, simple reordering rules are applied, such as moving nouns after verbs when translating from English to Arabic as shown in Figure 1. The disadvantages of using the direct approach include its inability to deal with phrases and larger structures, and to correctly handle longer distance reordering.

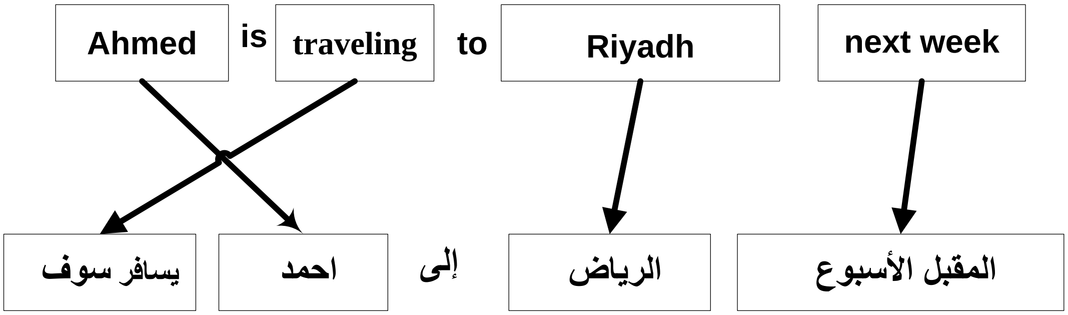


Figure 1: English to Arabic direct translation

In the transfer approach, the input text is first parsed into the source language parse structure, similar to the one shown in Figure 2. Then, reordering rules are applied to transform the source language parse structure into the target language parse structure. The target language sentence is then generated from the parse structure. Thus, machine translation in this approach involves three phases: analysis, transfer, and generation [2].

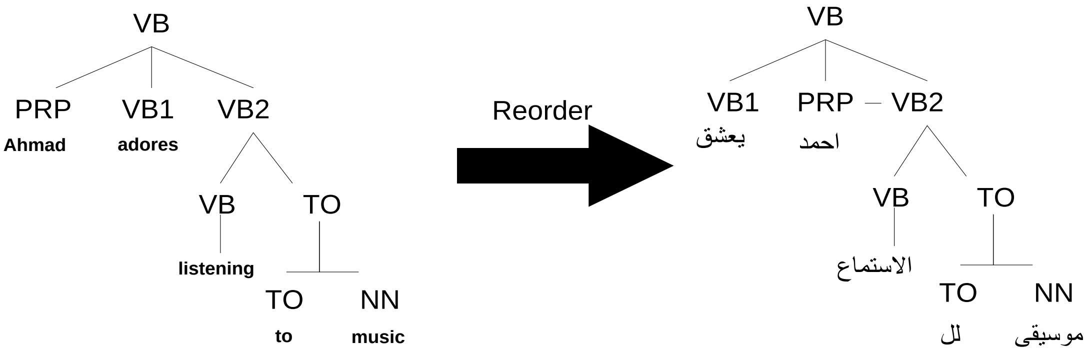


Figure 2: English to Arabic pars tree reordering

In the Interlingua approach, the source language text is analyzed and converted into some abstract meaning representation, called an Interlingua. The target language is then generated from this Interlingua representation. This approach is popular in many-to-many multilingual translation systems like that of the European Union languages [2].

SMT approaches utilize probabilistic models of faithfulness and fluency in the translation process. Faithfulness is defined as the proper translation of the source language that fully preserves the meaning. Fluency is the appropriateness of the translation in the destination language, in terms of its eloquence. These models are combined to choose the most probable translation. Using the product of faithfulness and fluency as a quality metric, one can model the translation from a source language sentence S to the best translation sentence $\hat{T}$ in the target language:

$$\hat{T} = \underset{T}{\operatorname{argmax}} \left(faithfullness(T, S) * fluency(T)\right) \qquad (1)$$

All SMT approaches are based on the idea of word alignment. A word alignment is a mapping between the source words and the target words in a set of parallel sentences, as shown in Figure 3. In those sentences, it is assumed that each sentence in a language is already known to correspond to which sentence in the other language [2].

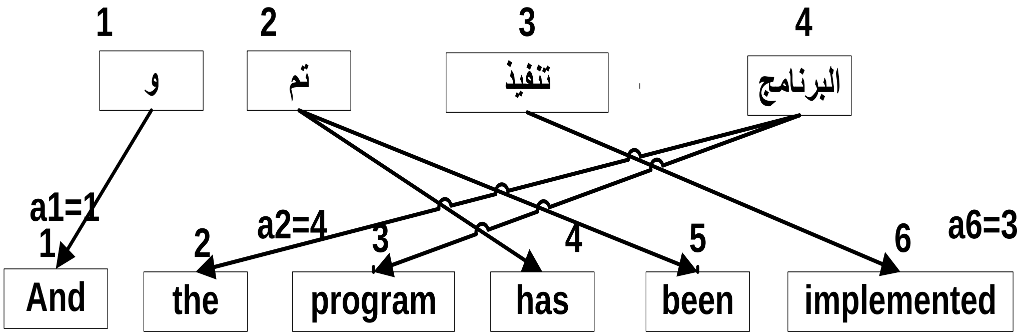


Figure 3: Arabic Text Aligns With English Text

1.2. Objective

The main objective of this thesis is to promote and advance research in English to Arabic machine translation. We build an English to Arabic machine translation system using the statistical phrase-based approach. We developed the Arabic language model from Watan and Khaleej corpora, in addition toa newly collected Arabic corpus that was built from different sources. In addition, three different parallel corpora (two readily-available corpora and another one that we developed) were used and their performance was analyzed. The following steps outline the thesis work:

1) Survey of different English to Arabic translation techniques; Classical MT approaches and Statistical MT approaches (SMT).

2) Develop an English to Arabic machine translation prototype system using the phrase based SMT. This is composed of the following tasks:

   i) Development of software that automatically collects Arabic natural language data.
   ii) Development of an Arabic monolingual corpus from the collected material in (i).
   iii) Development of the Arabic language model for the monolingual corpus.
   iv) Development of a sizable Arabic to English parallel corpus from Meedan news Website.
   v) Development of a phrase table using at least two Arabic to English parallel corpora (Meedan, UN and LDC).

3) Evaluation of the developed system using an automatic evaluation approach called the bilingual evaluation understudy (BLEU).

1.3. Contributions

The main contributions of this thesis include the following.

1 Developing an English-to-Arabic machine translation prototype system based on the phrase-based statistical machine translation approach [3].

2 Constructing an open source parallel Arabic–English corpus that is ready to use for automatic machine translation and a sizable open source Arabic monolingual corpus.

3 Evaluation of the developed system using the Bi-Lingual Evaluation Understudy (BLEU) approach.

1.4. Thesis Organization

The remainder of this thesis consists of the following chapters: Chapter 2 provides literature review of English to Arabic machine translation by presenting different research directions addressing this problem. It also outlines certain strengths and weaknesses for each approach. Chapter 3 presents the details of Statistical Machine Translation approaches with all their variants. We describe our proposed system for the system and the data sets used to training and testing in Chapter 4. In Chapter 5, we begin by discussing issues related to the performance evaluation of automatic machine translation systems followed by presenting the results obtained for various decoders and data corpuses, with explanations and justifications. Finally, we present our conclusions and suggested future work in Chapter 6.

CHAPTER 2

Literature Review

2.1. Introduction

Researchers in machine translation noticed that translation is not a word to word process. Knowledge of the surrounding words helps in generating a more accurate translation. In addition, one word in the source language may be translated into more than one word in the target language. This led researchers to investigate statistical machine translation methods that take the aforementioned issues into account. The first work on statistical machine translation was reported by Brown et al [4]. Advances in SMT lead to the usage of phrases as translation units instead of words. This produced phrase-based SMT methods [5, 6]. We divide our literature survey into two categories: the first one is related to general English to Arabic machine translation approaches, which is presented in Section 2; the second one is related to phrase-based statistical machine translation, which is presented in Section 3.

2.2. General English–Arabic Machine Translation

Early works in English to Arabic MT were largely based on the transfer classical MT approach [7-13] . Ibrahim et al [9] developed an English-to-Arabic translation system for embedded idioms and proverb expressions. Pease et al [8] developed a system which translates medical texts from English and German to Arabic. El-Desouki et al [7] used the prolog language to build an expert system for English-to-Arabic machine translation. Mokhtar et al [10] developed an English-to-Arabic MT system, which operates on abstracts from the field of Artificial Intelligence.

Al-Dam, et al. [13] developed a neural network-based English-to-Arabic Machine Translation System using an English-to-Arabic Bilingual Corpus from an unrestricted domain. In their approach, the transfer module employs neural networks to learn correspondences between source and target language structures using a large set of English sentences and their Arabic translations. They built a bilingual corpus from two Arabic books. The first one is titled "The Art of Translation From English to Arabic" [14], and the second one is titled "Your Guide to Correct Translation" [15]. The total size of the corpus is 2941 KB, with 23974 English word and 29679 Arabic words. Their system consists of three phases: The analysis phase, the neural-network-based transfer phase and the generation phase. The analysis phase deals with the lexical and syntactic analysis and produces the internal representation of the source language. The neural-network-based transfer module is an intermediary phase or between the analysis and the generation phases. The purpose of this module is to convert the source language (intermediate) representations into target language (intermediate) representations. All aspects of lexical or structural translational differences between the source and target languages are captured in this phase. The generation phase outputs the final translation in a form which is lexically and syntactically acceptable. This approach resulted in 81 perfectly translated sentences out of 200 test sentences.

Mohammed et al. [16] proposed an English to Arabic Machine Translation system that is based on the Reordering Algorithm that uses existing context-free grammar (CFG) format to identify the Part Of Speech (POS) for single words and reordering the CFG. It employs an English dictionary to translate single words and identify their categories from their tags in order to produce a correct translation using the CFG rules. For example, if a word is categorized as feminine, the subsequent adjective is chosen to be feminine. The system consists of two main phases. The first phase breaks the English sentences into its components until reaches the word level, generating suitable grammatical tags to each word. The second phase associates one Arabic meaning for each English word and aligns the target language words according to the target language rules. The domain area includes twenty abstracts containing ninety five sentences from the European Psychiatry Journal. This system achieved 81.8% according to the authors.

A hybrid approach combining the advantages of rule-based machine translation (RBMT) with the advantages of example-based machine translation (EBMT) was proposed by Alawneh and Sembok [17], The OAK Parser was used to analyze the input English text to get the part of speech (POS) for each word as a pre-translation process using the C# language. Validation rules for English and Arabic have been incorporated at the database level and the program level. The rules were classified into grammar rules, English-Arabic rules, linguistic rules and translation Rules. In RBMT, the process of conversion was based on the use of bilingual dictionaries and rules for converting source language structures into target language structures. The EBMT system extracted examples of target language sentences that are analogous to input source language sentences. The extraction of appropriate translated sentences was preceded by an analysis stage for the decomposition of input sentences into appropriate fragments. In hybrid MT, when an example of the source language to be translated into the target language is not found in the machine database, rules such as those in RBMT are used.

Translating English noun phrase (NP) into Arabic is as important as sentence translation, since NPs form the majority of textual content of the scientific and technical documents, as reported by Shaalan et al. [11]. They developed their system using SICStus Prolog and the parser was written in DCG formalism, where the DCG translates grammar rules directly into Prolog. To train their system, they collected 116 real titles of theses from the computer science domain. The training sample consisted of 50 titles out of the 116 titles, and the remaining 66 titles were used for evaluating their approach. They reported 92% accuracy as compared to a human translation.

A bi-directional English-Arabic machine translation system specifically developed for an expert system in the agriculture domain was proposed by Shaalanet al. using the transfer approach [12]. A set of real parallel 100 phrases and sentences from both English and Arabic from agricultural expert systems at CLAES was used to evaluate their approach. Overall BLEU evaluation scores of 0.6427 for translating from English to Arabic and 0.8122 for translating from Arabic to English were achieved.

2.3. Phrase-based Statistical Machine Translation

More recent research work in MT has shown that the phrase-based SMT gives more accurate translation results [18]. Due to the availability of English to Arabic parallel corpora and the lack of Arabic to English parallel corpora, many research works focused on Arabic to English SMT. Recently, there is an increase in research publications with respect to English to Arabic SMT.

Ahmed El Kholy et al [19] explored morphological tokenization schemes and orthographic normalization options for English to Arabic SMT. They use an English-Arabic parallel corpus of about 142K sentences from the Linguistic Data Consortium (LDC). The parallel text is mainly composed of Arabic News. The evaluation of the system obtained a BLEU score of 0.62. Hassan Al-Haj et al [20] addressed the challenge of translating English into Arabic, which is a morphologically rich language, using a phrase based statistical machine translation approach. They explored the full spectrum of Arabic segmentation schemes ranging from full word form to fully segmented forms, and examined their effects on system performance. They used 5 million sentence pairs from the LDC data corpus.

Syntactic reordering within an English to Arabic SMT system was proposed by Nizar Habash et al [21]. They achieved an increase over Free Pharaoh of 1.6% BLEU, which is a significant improvement in the translation. For training the system, they used English-Arabic parallel corpus consisting of 126K sentences extracted from the LDC corpus. Mona Diab et al [22] study the impact of Arabic diacritization on statistical machine translation by defining several diacritization schemes ranging from full to partial diacritization. The results showed that none of the partial or full diacritization schemes significantly enhanced performance over the non-diacritized baseline. On the other hand, a full diacritization scheme performed significantly worse than no diacritization. The training data consisted of about 5 million words of Arabic-English parallel news wire from the LDC corpus.

2.4. A summary of recent Literature Review

Table 1:A summary of literature review in English to Arabic machine translation.

Reference	Domain	Size of data	approach	Performance
[Rasha Al-Dam, Ahmed Guessoum , 2010] [13]	Unrestricted Domain	2941 KB	neural network-based (transfer)	80% human evaluation
(Mohammed and Ab Aziz , 2011) [16]	abstract from the European Psychiatry Journal	(20) abstracts containing (95) sentences	Reordering Algorithm using Context Free Grammar (CFG)	81.855% human evaluation
[Alawneh and Sembok , 2011][17]	general	--	hybrid-based	--
[Shaalan, Rafea et al. ,2004][11]	computer science	116 real titles of thesis	Transfer -prolog	92%
[Shaalan, K., A. Hendam, et al. ,2010] [12]	agriculture domain	100phrases and sentences	Transfer	0.6427 BLEU
[Ahmed El Kholy et al [19]	News	142K sentences Non free corpus from LDC	Orthographic and Morphological Processing phrase-based SMT	0.62 BLEU
[Nizar Habash et al,] [21]	News	126K sentences Non free corpus from LDC	Syntactic reordering phrase-based SMT	1.6 %BLEU increase over baseline system
[Mona Diab et al ,2007][22]	News	5 million words Non free corpus from LDC	Arabic diacritization phrase-based SMT	0.4195 BLEU

Table 1 outlines a summary of the literature review related to English to Arabic machine translation. One notices that the domain in most non-statistical based machine translation work is generally very restricted and small, with the exception of Al-Dam et al.'s work. The reason is that the transfer-based machine translation does not require huge data corpus to carry out the translation. In addition, since the domain is usually very restricted, the vocabulary, in turn, is not that huge. With respect to accuracy, we cannot compare the reported results as they do not come from the same queries. However, we can safely conclude that statistical machine translation approaches perform better than transfer-based approaches [18].

2.1. Available English To Arabic Translation sites and packages

The most popular English to Arabic machine translation systems include Google, Bing, yahoo, Al-Wafi, Al Mutarjim Al Kafi translators[23]. Google Inc. provides free online automatic machine translation system which is based on statistical approach [24]. Google translate uses hundreds of millions of documents that have previously been translated by human translators to build the translation model, which currently supports up to 58 languages [24]. Bing translator was developed by Microsoft research department , just like Google translate, this system uses statistical machine translation approached [25]. However, currently it translates only 32 languages Arabic inclusive [26]. Yahoo Inc. uses a commercial machine translation system for its online automatic machine translation which is provided by SYSTRAN Company, this system implement a hybrid approach combining the rule-based machine translation with statistical machine translation approaches. All previously mention systems are free online translation for multiple languages including Arabic. However, Al-Wafi, al Mutarjim Al Kafi translators are commercially specialized English/Arabic machine translators designed for ordinary user [27].

CHAPTER 3

Statistical Machine Translation

3.1. Introduction

Traditional MT architectures are used to understand the representation of the source language in order to perform the translation to the target language. they, also, focus on a procedure that makes it feasible to translate the source language to the target language. Another way to tackle the problem of translation is to focus on the desired result, not the translation process. It could be noted that the consensus in translating a single sentence seems impossible for a sentence in one language to be translated to a sentence in another one, in the strict sense of the word (semantic meaning). For example translating the English phrase " put yourself in my shoes " into Arabic as " ضع نفسك في حذائي " is not appropriate, compared to the semantic proverbial translation: " ضع نفسك في مكاني ". Each language has culture-specific concepts, a metaphor, a construction and a word, or a tense without an exact parallel in the other language[2]. The desired translation we need is the one that is faithful to the source language, not changing its intended meaning, and as native to the target language. However, achieving such a translation automatically is close to impossible, most of the time [17, 28].

When a translator translates text from one language to another, he, first, fully understands the text written in the source language. Then, he evokes all vocabulary that can be used to carry out the translation. Finally, he decides to exclude all but one word or set of words that he deems best in the translation [2]. This provides the basis on which the statistical machine translation approach is built upon. In particular, the problem of machine translation can be modeled as a maximization problem, by creating probabilistic models of faithfulness and fluency, and then combining these two models to choose the best-translation. Therefore, the problem of translating an English language sentence E to an Arabic language sentence A can be modeled as [2]:

$$\hat{A} = argmax_A\, (faithfulness\,(A, E) \times fluency\,(A)) \qquad \textbf{(2)}$$

Where $E = e_1, e_2, \ldots, e_m$ consists of m English words and $\hat{A} = a_1, a_2, \ldots, a_l$ is the best Arabic sentence corresponding to E with highest probability $P(A|E)$.

$$\hat{A} = argmax_A\, P(A|E) \qquad \textbf{(3)}$$

Using Bayes rule, Equation (3) can be rewritten as:

$$\hat{A} = \frac{argmax_A \, P(E|A) \cdot P(A)}{P(E)} \qquad \textbf{(4)}$$

Since we are choosing the best Arabic sentence for a fixed English sentence E, and since $P(E)$ is a constant, we can ignore $P(E)$ inside the argmax, and hence Equation (4) becomes.

$$\hat{A} = argmax_A \, P(E|A) \cdot P(A) \qquad (5)$$

Equation (5) remodels the problem by making it investigate "all sentences" in Arabic and then choosing the best one corresponding to Sentence E. Such noisy channel representation shows that finding the best translation requires two components: a translation model $P(E|A)$, and a language model $P(A)$ [2]. In order to generate such translation, a third component is needed, which is called the decoder.

Research work in statistical machine translation can be classified into two categories [2]: word-based and phrase-based. Sections 2 and 3 in this chapter elaborate more on these approaches.

3.2. Word-Based Models

Word-based models consider words as the atomic units for translation. As a result, existence of multiple synonyms poses a challenge to automatic translation, especially when the target language is a morphologically rich language, which is the case for Arabic. For example, translation of the English word "house" into Arabic can be one of the following words منزل ، بيت ، دار ،مسكن ، عائلة ، مأوى ، أسرة ، مجلس تشريعي Some of these are commonly used such as منزل ، بيت ، دار ، مسكن while others are rarely used [18].

Therefore, word-based statistical machine translation approaches employ word usage statistics in the parallel corpus, in addition to the language model in choosing the best translation. For example, assume that we have an English to Arabic parallel corpus that needs to translate the word "house". The translation is carried out by finding the different Arabic words used in translating "house" in addition to their frequency. Table 2 shows hypothetical frequencies for the aforementioned listed translations of the word "house".

Table 2: Hypothetical counts of translations of the English word house into Arabic.

Arabic Translation	Count	Arabic Translation	Count
منزل	700	عائلة	50
بيت	500	مأوى	60
دار	300	أسرة	50
مسكن	207	مجلس تشريعي	3

The straightforward way to compute the probability distribution for a given Arabic translation of an English word in Table 2 is to use the ratio of the number of occurrences of that Arabic word over the total number of all possible Arabic translations. Since we have 1870 occurrences of the word house in our text collection, 700 of which have been translated as منزل, we have *p*(*house* [منزل]) = 0.37. Figure 4 shows the probability of translation for each choice and results in the word منزل being the most common translation of the word house.

$$p_{house}(a) = \begin{cases} 0.374 & \text{if a= منزل} \\ 0.267 & \text{if a= بيت} \\ 0.160 & \text{if a= دار} \\ 0.111 & \text{if a= مسكن} \\ 0.027 & \text{if a= عائلة} \\ 0.032 & \text{if a= مأوى} \\ 0.027 & \text{if a= أسرة} \\ 0.002 & \text{if a= مجلس تشريعي} \end{cases}$$

Figure 4: a probability distribution for all eight choices of translations of the word "house".

This way of estimating the probability distribution from data is called maximum likelihood estimation, as it maximizes the likelihood of data. There are many ways to build a model for the given data [18]. However, in a general domain, word based approaches tend to be biased toward highest frequency translations without any reference to the context in which the word appears in.

3.3. Phrase-Based Models

Phrase based models are among the most successful models of statistical machine translation that have gained more popularity in recent years. This model uses the phrase as the atomic unit of translation, instead of the word. Hence, probabilities are estimated based on phrases (sequences of words) as well as single words. To carry out the translation, a parallel corpus consisting of sentences and their translations is used. Entire phrases often need to be translated and moved as a unit as shown in Figure 5 [2].

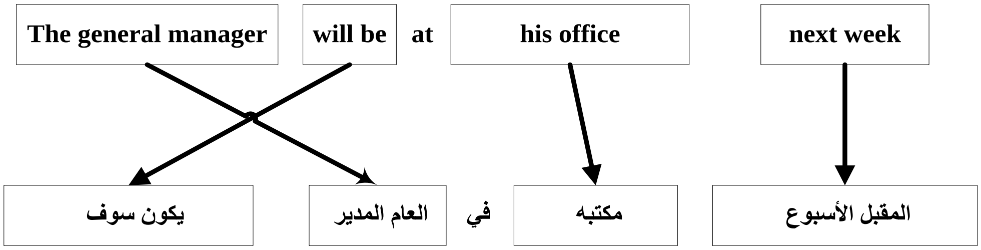


Figure 5: phrase reordering when translating from English to Arabic .

In the phrase-based SMT approach, we first group the English source words into phrases e_1, e_2... e_I. Next we use a probability model to translate each English phrase e_i into an Arabic phrase a_j. This process is called the decoding process. The probability model for phrase-based translation depends on the translation model probability and the language model probability [2]. A parallel corpus, like the one shown in Figure 6, is used as training data to build this probability model. Finally, reordering of each Arabic phrase is carried out, if necessary. The following sections describe in more details the decoding process and the translation and language models, which are essential components of the phrase-based SMT approach.

مكنز ثنائي Parallel corpus (Arabic → English)		
Saudi Arabia: Euro money 2010 exposes the reality of global financial markets		**السعودية: "يورومني 2010" يناقش واقع الأسواق المالية العالمية**
The fifth round of Riyadh’s Euro money 2010 festival began today. Global markets have been witnessing some instability in light of the Greek financial crisis.		انطلقت اليوم فعاليات النسخة الخامسة لمؤتمر "يورومني 2010" بالرياض، في خضم عدم استقرار تشهده مجموعة من الأسواق العالمية على خلفية الأزمة المالية اليونانية.
This round has witnessed the attendance of more than 1,200 financial and business figures within the region and from different areas of the world, in an effort to study the reality of financial markets and the impact of global crises on them.		وتتميز هذه الدورة بحضور أكثر من 1200 شخصية من عالم المال والأعمال من المنطقة ومن مختلف دول العالم من أجل تدارس واقع الأسواق المالية في العالم وتأثير الأزمات الاقتصادية عليها.
The conference will be held over two days, including eight main sessions and a number of workshops run by global financial institutions		وتمتد أعمال المؤتمر على مدى يومين، تشمل ثماني جلسات رئيسية ومجموعة من الورشات التي ستديرها بعض المؤسسات المالية العالمية.

Figure 6: Parallel corpus (Arabic - English) [29]

A. *Decoding*

Decoding is the process of finding a target translation sentence (Arabic) that best corresponds to a source sentence (English) using a translation model and a language model. Decoding is an NP-complete problem [30], as there is an exponential number of choices for the target sentence. Hence, MT decoders use heuristics to search for a target sentence that maximizes its probability based on the translation model and the language model. This search process starts with the null hypothesis as an initial search state. This hypothesis is expanded by choosing every possible English word or phrase that could generate Arabic sentences. The search process terminates when the best translation is reached [18].

B. *Language Model*

The language model is an essential part of the statistical machine translation system, which measures how a native speaker of the language would express a sequence of words. A language model is a function that takes a sentence as input and returns the probability of its occurrence in the language text. In order to build the language model, we need to collect a sizable monolingual corpus and count how many sequences of words occur in it. The most common method for building the language model is using N-gram language models. N-Gram language models use the Markov model to break the probability of a sentence into the product of the probability of each word; The N-gram measure is related to how much probability is given to a set of unigram, bigram, trigram and N-gram words. For example, calculating a trigram language model requires collecting the probability of occurrence of every three-word sequence from a large amount of text. Larger language models usually lead to better results in the translation process.

*C. **Translation model***

The translation model uses a source sentence S and a target sentence T to compute the probability P(T|S) which is the probability of S generating T. The parallel corpus is used to estimate these probabilities.

3.4. Evaluation of Translation Quality

In order to compare and contrast different automatic machine translation approaches, we need to have a suitable metric that judges the quality of the translation. One commonly used metric that assigns a value that reflects the quality of the translation is Bi-Lingual Evaluation Understudy approach (BLEU)[31]. BLEU compares the n-grams of the candidate translation with the n-grams of the reference translation, counting the number of matches. To compute the BLEU, the geometric mean of the modified precision scores, p_n, of the test corpus, shown in Equation (6), is multiplied by a brevity penalty factor, shown in Equation (7) [31]. The positive weights w_n in Equation (6) are used in the geometric average, and therefore sum to 1. The brevity penalty factor depends on the length of the candidate translation sentence, c, and the effective reference corpus sentence length, r. The BLEU value shown in Equation (8) ranges between 0 and 1.

$$GM = \exp\left(\sum_{1}^{\mathrm{N}} \mathrm{w_n} \log \mathrm{p_n}\right) \tag{6}$$

$$BP = \begin{cases} 1, & c > r \\ e^{(1-r/c)}, & c \leq 0 \end{cases} \tag{7}$$

$$\text{BLEU}=GM.BP \tag{8}$$

The aim of this research work is to develop an English-to-Arabic machine translation system, which implements the phrase-based statistical machine translation method using open source parallel corpora and to study and analyze its performance. The next chapter introduces our developed system for phrase-based MT from English to Arabic.

CHAPTER 4

Phrase-Based Arabic Machine Translation Framework

4.1. Introduction

In order to build a SMT system, we need various tools and training data. In particular, we need a parallel corpus, a monolingual corpus, word alignment software, a language modeling toolkit, and a decoder software. The overall phrase based MT system is shown in Figure 7. First, the source text (English text) that needs to be translated to its corresponding Arabic text is input into the system. Second, we carry out the preprocessing step, which includes tokenization, conversion of the text into lowercase and removal of unwanted words. Tokenization involves removal of extra spaces and the separation of words from punctuation marks and special symbols like the dollar sign. Third, the decoding task is performed where we determine the best translation. Finally, the post-processing stage is carried out in order to produce the final result (the target Arabic sentence) to the end user. Details of each stage are included in the subsequent sections.

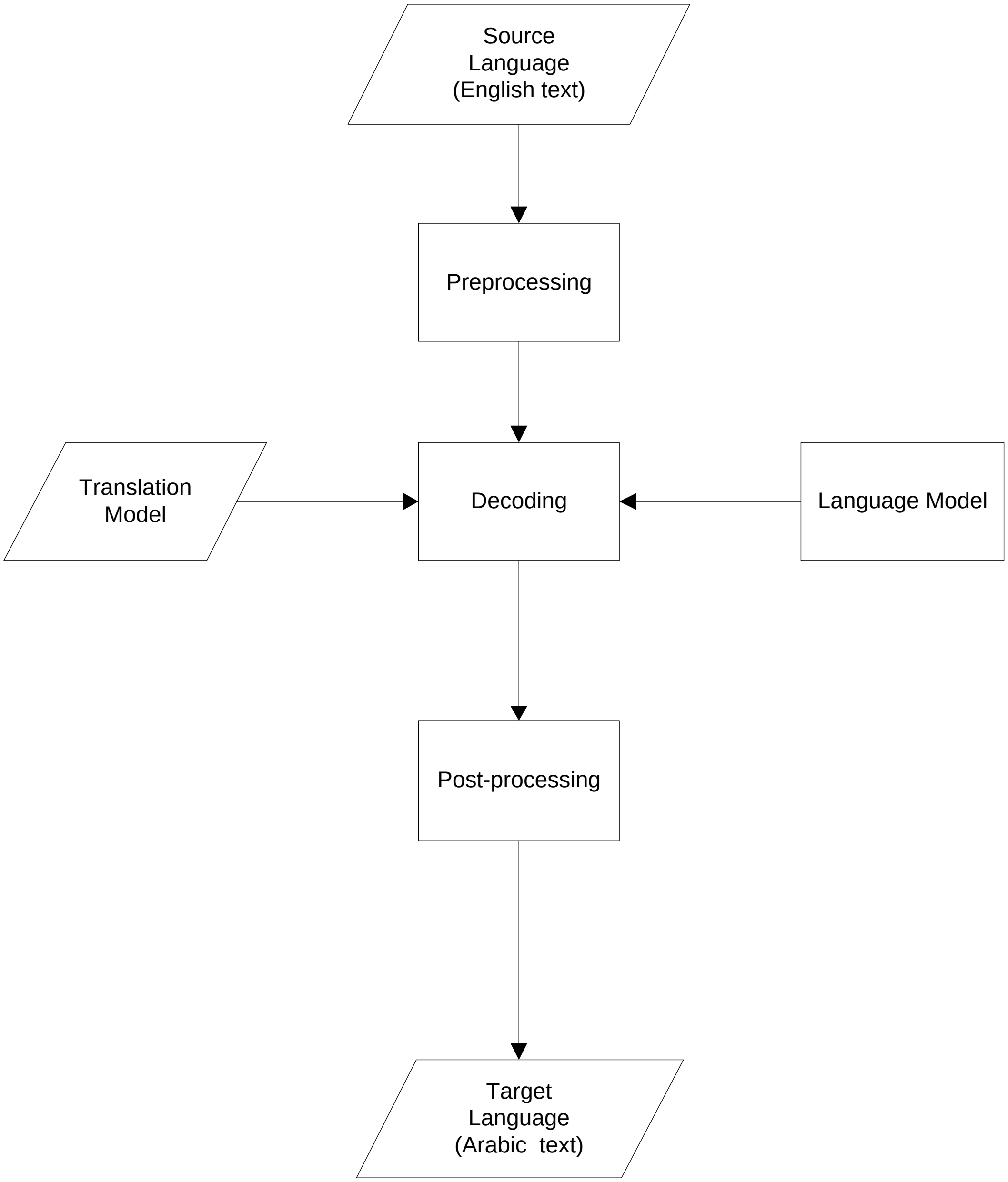


Figure 7: Architecture of the Proposed System.

4.2. Corpus Development

A corpus or a text corpus in the field of linguistics and natural language processing (NLP) is a large and organized set of texts, usually electronically stored and processed [32]. Phrase-based MT systems employ two types of corpora: a monolingual corpus and a parallel corpus. A monolingual corpus is a corpus that contains texts in a single language. A parallel corpus is a collection of texts, each of which is translated into one or more than one language [33, 34]. The simplest case of a parallel corpus is built for two languages, as shown in Figure 6. SMT systems utilize available linguistic data, translated by human efforts, in their development. The language model and the translation model act as input data to the system. To train the language model, we need to collect a sizable Arabic monolingual corpus of the target language (Arabic). As for the translation model, we need to build an Arabic to English parallel corpus that will be used to train the translation system. To build this parallel corpus, a large volume of high quality Arabic to English human translated text needs to be collected. The following sections describe our efforts in building both corpora.

A. *Monolingual Corpus Development*

In order to build the monolingual Arabic corpus, two freely available Arabic sources on the internet were utilized, viz., Watan and Khaleej [35]. Watan-2004 corpus contains about 20000 html articles spanning six topics: culture, religion, economy, local news, international news and sports. Khaleej-2004 corpus contains html articles in the topics of international news, local news, economy and sports [35]. In addition, The available Arabic text from Meedan website and the UN data have been utilized in building the corpus [29, 36, 37]. Table 3 outlines the number of documents in each corpus.

Table 3: Corpora Topics And Number Of Documents For Each Topics[38] .

Topic	Corpus Size (Number of documents)	# of word
Khaleej-2004 corpus		
International News	953	534532
Local News	2398	996205
Economy	909	418978
Sports	1430	551728
Total number of docs	5690	2501443
Watan-2004 corpus		
Culture	2782	1406402
Religion	3860	3138623
Economy	3468	1470099
Local News	3596	1562042
International News	2035	862472
Sports	4550	1437148
Total number of docs	20291	9876786
Additional freely-available articles		
Meedan	19956 (sentences)	426863
UN	72338 (sentences)	746679

All these documents were processed and saved as one file using Unicode format in order to be used by the language model tool. The total number of distinct words in the corpus is equal to 402964 words.

B. *Parallel Corpus Development*

SMT requires a huge amount of parallel text in the source and target language in order to achieve high quality translation. Arabic language electronic content availability is classified as low, compared to other languages' content. In fact, although Arabic is ranked at the top five languages spoken on Earth, its content ranks tenth, occupying less than 2% of electronic content [39]. Therefore, developing a sizable parallel corpus from freely available sources posed a great challenge in this thesis work.In addition, the collected documents need to be preprocessed prior to their use in the SMT system. Hence, we developed our own parallel corpus from Meedan news website [29]. Meedan is a nonprofit cross-cultural forum for disseminating and discussing news in Arabic and English [29]. The Meedan website distributes Arabic news along with their English translation in html format, as shown Figure 8. For the purposes of evaluating the suitability of our corpus, we used two more parallel corpora that are available to us: viz., the parallel corpus developed by LDC [37], which is proprietary, and the United Nations corpus [36], which is a free corpus available on the Internet. The LDC corpus was developed by Xiaoyi Ma and Dalal Zakhary at 2007 under a research project titled “Arabic Broadcast News Parallel Text - Part 1”. The United Nations corpus contains document collections from the UN that have been compiled for research purposes.

The currently available documents in their original form are not ready to be used in the SMT system. Therefore, we first need to preprocess these documents by converting them into a format that is suitable for the system. Figure 9 shows a sample ready-to-use format of the parallel corpus. To achieve this, a program written in MS Visual Studio

2010 has been developed to extract the English and the Arabic contents from the various sources and to generate the ready-to-process parallel corpus. Correctly-aligned Arabic-English sentence pairs have been extracted from all corpora data and have been subjected to our document cleaning preprocessing step. Our document cleaning step consists of removing the following categories of sentences:

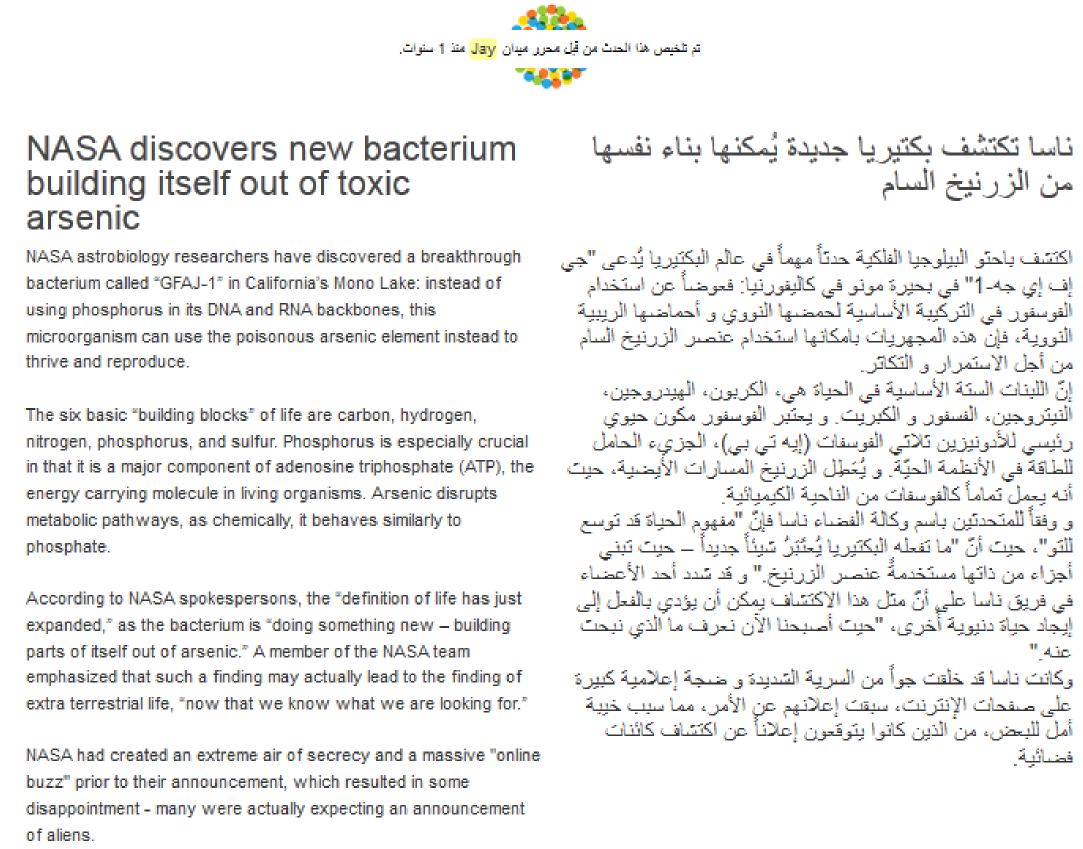
- Very long sentences. These sentences are removed due to the fact that GIZA++ tool, which is detailed in Section 4.4, suffers from performance degradation when the number of characters exceeds 200 characters.
- Sentences with mixed English and Arabic words. Having mixed words in English or in Arabic most probably constitutes a sign of the existence of transcription mistakes, rendering the Arabic-English sentence pair to be incorrect. In addition, even in the absence of transcription mistakes, the existence of mixed words in one sentence may greatly reduce the quality of the translation.
- Arabic-English sentence pairs, in which the difference between the number of characters in English and the number of characters in the corresponding Arabic translation exceeds 70 characters. The reason for this is that it could, very well, be a sign of translation by summarization, in which not all or almost all words are translated.

With respect to Meedan documents, they are html files containing both the Arabic and their corresponding English translations. The program converts each html file into a Unicode text file, removing all html tags before the cleaning process is carried out.

Special preprocessing was neither needed for LDC data nor for UN data before the cleaning process. Finally, the Arabic-English sentence pairs are written into two files, one for English sentences and the other for Arabic sentences, preserving their order. Table 4 shows the corpora sizes before and after the document cleaning process. A snapshot of the developed program is shown in Figure 10.

Table 4: Parallel corpus size before and after cleaning.

Corpus	Original Size (# sentences)	Size After Cleaning (# sentences)
LDC	4901	3779
Meedan	19956	17863
Un	72338	48499



تم تلخيص هذا الحدث من قبل محرر ميدان Jay منذ 1 سنوات.

NASA discovers new bacterium building itself out of toxic arsenic

NASA astrobiology researchers have discovered a breakthrough bacterium called "GFAJ-1" in California's Mono Lake: instead of using phosphorus in its DNA and RNA backbones, this microorganism can use the poisonous arsenic element instead to thrive and reproduce.

The six basic "building blocks" of life are carbon, hydrogen, nitrogen, phosphorus, and sulfur. Phosphorus is especially crucial in that it is a major component of adenosine triphosphate (ATP), the energy carrying molecule in living organisms. Arsenic disrupts metabolic pathways, as chemically, it behaves similarly to phosphate.

According to NASA spokespersons, the "definition of life has just expanded," as the bacterium is "doing something new – building parts of itself out of arsenic." A member of the NASA team emphasized that such a finding may actually lead to the finding of extra terrestrial life, "now that we know what we are looking for."

NASA had created an extreme air of secrecy and a massive "online buzz" prior to their announcement, which resulted in some disappointment - many were actually expecting an announcement of aliens.

ناسا تكتشف بكتيريا جديدة يُمكنها بناء نفسها من الزرنيخ السام

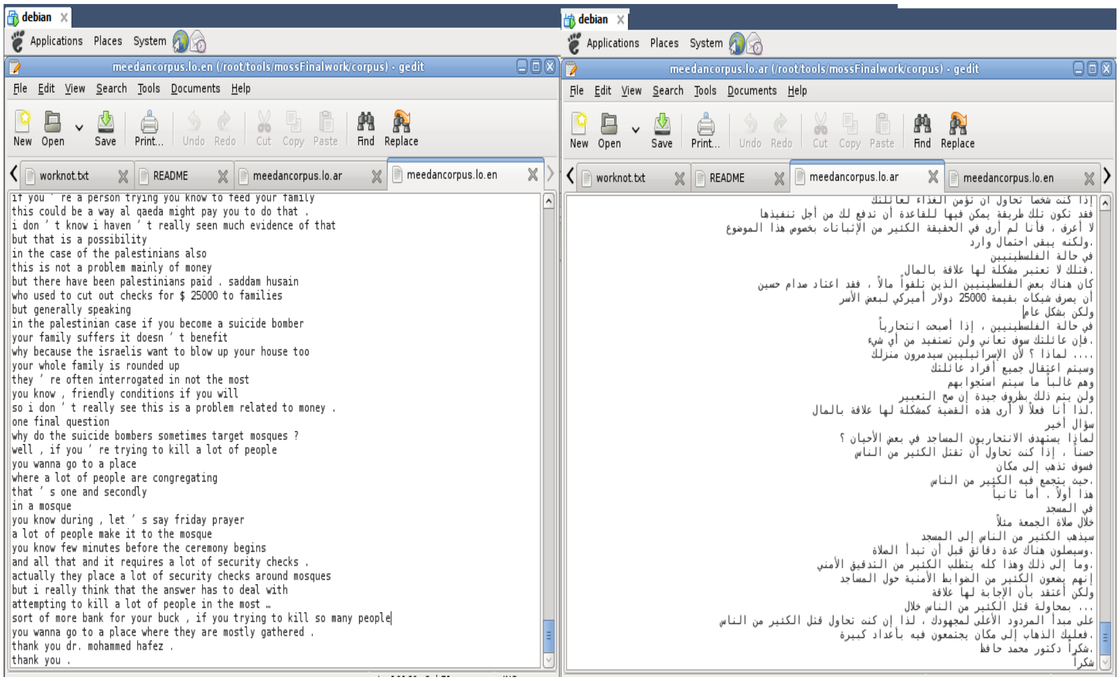
اكتشف باحثو البيلوجيا الفلكية حدثاً مهماً في عالم البكتيريا يُدعى "جي إف إي جه-1" في بحيرة مونو في كاليفورنيا: فعوضاً عن استخدام الفوسفور في التركيبة الأساسية لحمضها النووي و أحماضها الريبية النووية، فإن هذه المجهريات بامكانها استخدام عنصر الزرنيخ السام من أجل الاستمرار و التكاثر.

إنّ اللبنات الستة الأساسية في الحياة هي، الكربون، الهيدروجين، النيتروجين، الفسفور و الكبريت. و يعتبر الفوسفور مكون حيوي رئيسي للأدونيزين ثلاثي الفوسفات (إيه تي بي)، الجزيء الحامل للطاقة في الأنظمة الحيّة. و يُعطِل الزرنيخ المسارات الأيضية، حيث أنه يعمل تماماً كالفوسفات من الناحية الكيميائية.

و وفقاً للمتحدثين باسم وكالة الفضاء ناسا فإنّ "مفهوم الحياة قد توسع للتو"، حيث أنّ "ما تفعله البكتيريا يُعتَبَرُ شيئاً جديداً – حيث تبني أجزاء من ذاتها مستخدمةً عنصر الزرنيخ." و قد شدد أحد الأعضاء في فريق ناسا على أنّ مثل هذا الاكتشاف يمكن أن يؤدي بالفعل إلى إيجاد حياة دنيوية أخرى، "حيث أصبحنا الآن نعرف ما الذي نبحث عنه."

وكانت ناسا قد خلقت جواً من السرية الشديدة و ضجة إعلامية كبيرة على صفحات الإنترنت، سبقت إعلانهم عن الأمر، مما سبب خيبة أمل للبعض، من الذين كانوا يتوقعون إعلاناً عن اكتشاف كائنات فضائية.

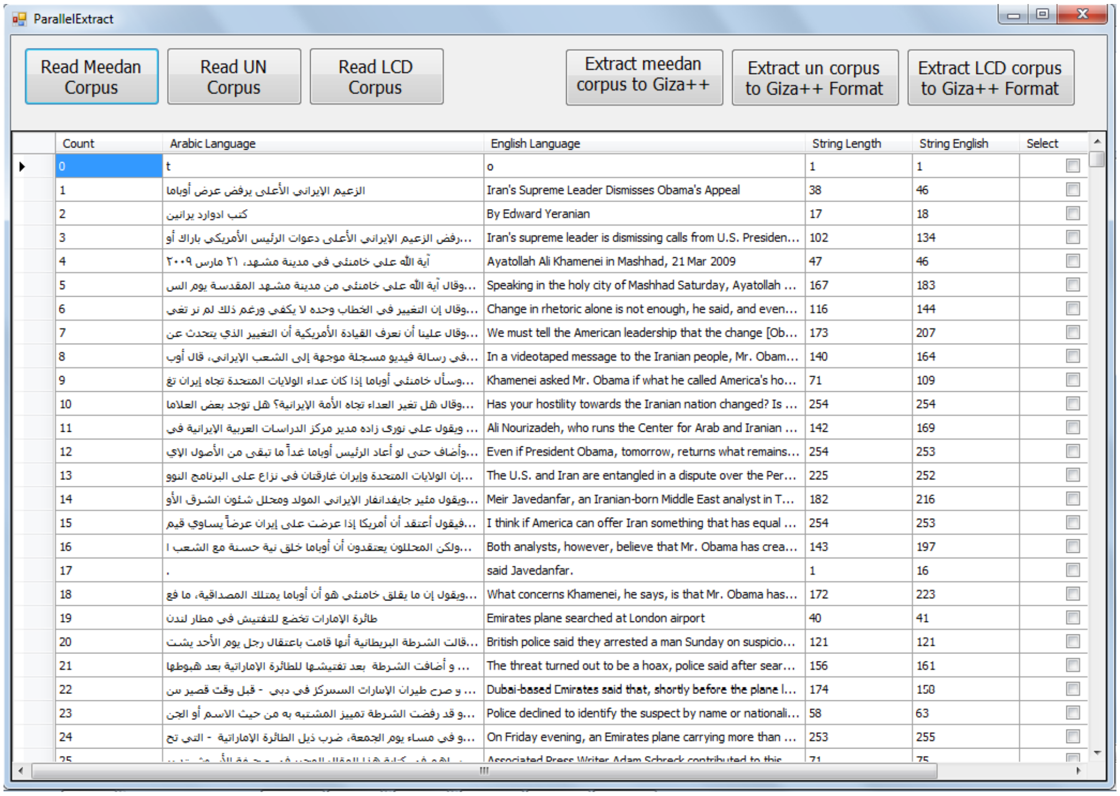
Figure 8: Meedan website The narrative of the original Arabic and the translated English.


debian
Applications Places System
meedancorpus.lo.en (/root/tools/mossFinalwork/corpus) - gedit
File Edit View Search Tools Documents Help
New Open Save Print... Undo Redo Cut Copy Paste Find Replace
worknot.txt README meedancorpus.lo.ar meedancorpus.lo.en
debian
Applications Places System
meedancorpus.lo.ar (/root/tools/mossFinalwork/corpus) - gedit
File Edit View Search Tools Documents Help
New Open Save Print... Undo Redo Cut Copy Paste Find Replace
worknot.txt README meedancorpus.lo.ar meedancorpus.lo.en

```
if you ’ re a person trying you know to feed your family
this could be a way al qaeda might pay you to do that .
i don ’ t know i haven ’ t really seen much evidence of that
but that is a possibility
in the case of the palestinians also
this is not a problem mainly of money
but there have been palestinians paid . saddam husain
who used to cut out checks for $ 25000 to families
but generally speaking
in the palestinian case if you become a suicide bomber
your family suffers it doesn ’ t benefit
why because the israelis want to blow up your house too
your whole family is rounded up
they ’ re often interrogated in not the most
you know , friendly conditions if you will
so i don ’ t really see this is a problem related to money .
one final question
why do the suicide bombers sometimes target mosques ?
well , if you ’ re trying to kill a lot of people
you wanna go to a place
where a lot of people are congregating
that ’ s one and secondly
in a mosque
you know during , let ’ s say friday prayer
a lot of people make it to the mosque
you know few minutes before the ceremony begins
and all that and it requires a lot of security checks .
actually they place a lot of security checks around mosques
but i really think that the answer has to deal with
attempting to kill a lot of people in the most …
sort of more bank for your buck , if you trying to kill so many people
you wanna go to a place where they are mostly gathered .
thank you dr. mohammed hafez .
thank you .
```

```
إذا كنت شخصا تحاول أن تؤمن الغذاء لعائلتك
فقد تكون تلك طريقة يمكن فيها للقاعدة أن تدفع لك من أجل تنفيذها
لا أعرف ، فأنا لم أرى في الحقيقة الكثير من الإثباتات بخصوص هذا الموضوع
.ولكنه يبقى احتمال وارد
في حالة الفلسطينيين
.فتلك لا تعتبر مشكلة لها علاقة بالمال
كان هناك بعض الفلسطينيين الذين تلقوا مالاً ، فقد اعتاد صدام حسين
أن يصرف شيكات بقيمة 25000 دولار أميركي لبعض الأسر
ولكن بشكل عام
في حالة الفلسطينيين ، إذا أصبحت انتحارياً
.فإن عائلتك سوف تعاني ولن تستفيد من أي شيء
.... لماذا ؟ لأن الإسرائيليين سيدمرون منزلك
وسيتم اعتقال جميع أفراد عائلتك
وهم غالباً ما سيتم استجوابهم
ولن يتم ذلك بظروف جيدة إن صح التعبير
.لذا أنا فعلاً لا أرى هذه القضية كمشكلة لها علاقة بالمال
سؤال أخير
لماذا يستهدف الانتحاريون المساجد في بعض الأحيان ؟
حسناً ، إذا كنت تحاول أن تقتل الكثير من الناس
فسوف تذهب إلى مكان
.حيث يتجمع فيه الكثير من الناس
هذا أولاً . أما ثانياً
في المسجد
خلال صلاة الجمعة مثلاً
سيذهب الكثير من الناس إلى المسجد
.وسيصلون هناك عدة دقائق قبل أن تبدأ الصلاة
.وما إلى ذلك وهذا كله يتطلب الكثير من التدقيق الأمني
إنهم يضعون الكثير من الضوابط الأمنية حول المساجد
ولكن أعتقد بأن الإجابة لها علاقة
... بمحاولة قتل الكثير من الناس خلال
على مبدأ المردود الأعلى لمجهودك ، لذا إن كنت تحاول قتل الكثير من الناس
.فعليك الذهاب إلى مكان يجتمعون فيه بأعداد كبيرة
.شكراً دكتور محمد حافظ
شكراً
```

Figure 9: A sample format of an English file and its corresponding Arabic translation file that are ready to be used in the SMT system.


ParallelExtract
Read Meedan Corpus
Read UN Corpus
Read LCD Corpus
Extract meedan corpus to Giza++
Extract un corpus to Giza++ Format
Extract LCD corpus to Giza++ Format

Count	Arabic Language	English Language	String Length	String English	Select
0	t	o	1	1	□
1	الزعيم الإيراني الأعلى يرفض عرض أوباما	Iran's Supreme Leader Dismisses Obama's Appeal	38	46	□
2	كتب ادوارد يرانين	By Edward Yeranian	17	18	□
3	رفض الزعيم الإيراني الأعلى دعوات الرئيس الأمريكي باراك أو...	Iran's supreme leader is dismissing calls from U.S. Presiden...	102	134	□
4	آية الله علي خامنئي في مدينة مشهد، ٢١ مارس ٢٠٠٩	Ayatollah Ali Khamenei in Mashhad, 21 Mar 2009	47	46	□
5	وقال آية الله علي خامنئي من مدينة مشهد المقدسة يوم الس...	Speaking in the holy city of Mashhad Saturday, Ayatollah ...	167	183	□
6	وقال إن التغيير في الخطاب وحده لا يكفي ورغم ذلك لم نر تغي...	Change in rhetoric alone is not enough, he said, and even...	116	144	□
7	وقال علينا أن نعرف القيادة الأمريكية أن التغيير الذي يتحدث عن...	We must tell the American leadership that the change [Ob...	173	207	□
8	في رسالة فيديو مسجلة موجهة إلى الشعب الإيراني، قال أوب...	In a videotaped message to the Iranian people, Mr. Obam...	140	164	□
9	وسأل خامنئي أوباما إذا كان عداء الولايات المتحدة تجاه إيران تغ...	Khamenei asked Mr. Obama if what he called America's ho...	71	109	□
10	وقال هل تغير العداء تجاه الأمة الإيرانية؟ هل توجد بعض العلاما...	Has your hostility towards the Iranian nation changed? Is ...	254	254	□
11	ويقول علي نوري زاده مدير مركز الدراسات العربية الإيرانية في ...	Ali Nourizadeh, who runs the Center for Arab and Iranian ...	142	169	□
12	وأضاف حتى لو أعاد الرئيس أوباما غداً ما تبقى من الأصول الإي...	Even if President Obama, tomorrow, returns what remains...	254	253	□
13	إن الولايات المتحدة وإيران غارقتان في نزاع على البرنامج النوو...	The U.S. and Iran are entangled in a dispute over the Per...	225	252	□
14	ويقول مئير جايفدانفار الإيراني المولد ومحلل شئون الشرق الأو...	Meir Javedanfar, an Iranian-born Middle East analyst in T...	182	216	□
15	فيقول أعتقد أن أمريكا إذا عرضت على إيران عرضاً يساوي قيم...	I think if America can offer Iran something that has equal ...	254	253	□
16	ولكن المحللون يعتقدون أن أوباما خلق نية حسنة مع الشعب ا...	Both analysts, however, believe that Mr. Obama has crea...	143	197	□
17	.	said Javedanfar.	1	16	□
18	ويقول إن ما يقلق خامنئي هو أن أوباما يمتلك المصداقية، ما فع...	What concerns Khamenei, he says, is that Mr. Obama has...	172	223	□
19	طائرة الإمارات تخضع للتفتيش في مطار لندن	Emirates plane searched at London airport	40	41	□
20	قالت الشرطة البريطانية أنها قامت باعتقال رجل يوم الأحد يشت...	British police said they arrested a man Sunday on suspicio...	121	121	□
21	و أضافت الشرطة بعد تفتيشها للطائرة الإماراتية بعد هبوطها ...	The threat turned out to be a hoax, police said after sear...	156	161	□
22	و صرح طيران الإمارات السسركز في دبي - قبل وقت قصير من ...	Dubai-based Emirates said that, shortly before the plane l...	174	150	□
23	و قد رفضت الشرطة تمييز المشتبه به من حيث الاسم أو الجن...	Police declined to identify the suspect by name or nationali...	58	63	□
24	و في مساء يوم الجمعة، ضرب ذيل الطائرة الإماراتية - التي تح...	On Friday evening, an Emirates plane carrying more than ...	253	255	□
25	[illegible]	Associated Press Writer Adam Schreck contributed to this	71	75	□

Figure 10: Snapshot of the program that build the parallel corpus.

4.3. Language Model Development

Building the language model is based on the development of the Arabic monolingual corpus that will be used to generate it. Figure 11 outlines the steps involved in building the language model. The most common method for building the language model is using N-gram language models.

Many open source tools can be used to build the N-gram language model. One of the most commonly used tools is the one developed by SRI International's Speech Technology and Research Laboratory, which is called SRILM. SRILM is a statistical language model toolkit that can be utilized in machine translation systems [40]. Its development started in 1995 and it consists of a set of C++ class libraries implementing language models, supporting data structures and miscellaneous utility functions. SRILM runs on UNIX and Windows platforms. It can be downloaded under an open source community license [40]. However, due to its inefficiency in loading time and memory management, we resorted to another tool, the IRSTLM toolkit.

The IRSTLM language model toolkit is an open source toolkit that is distributed under sourceforge.net [41]. IRSTLM is suitable for estimating, storing and accessing very large language models [41]. It can efficiently handle language models with billions of N-grams on traditional computing power. This toolkit can also employ parallel processing by distributing N-gram gathering and smoothing over a cluster of computers. In addition, it performs language model compression through probability quantization and lazy-loading of huge language models from secondary storage. It was integrated to work with both Moses and Pharaoh decoders. We were able to generate the 5-gram language model for our monolingual corpus. An example segment of the output language model in the forms of unigram, bigram and trigram is shown in Figure 12.

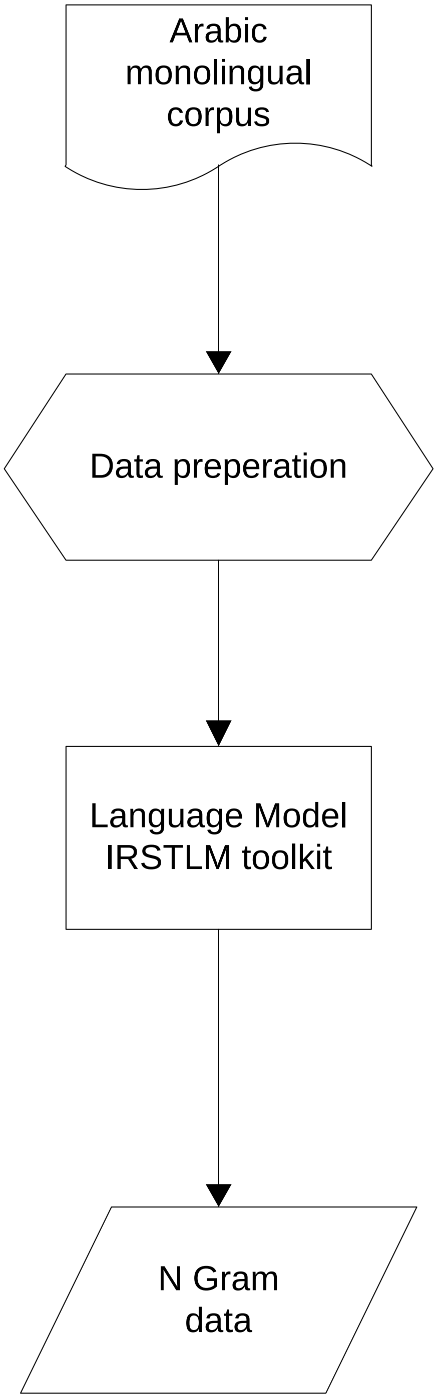
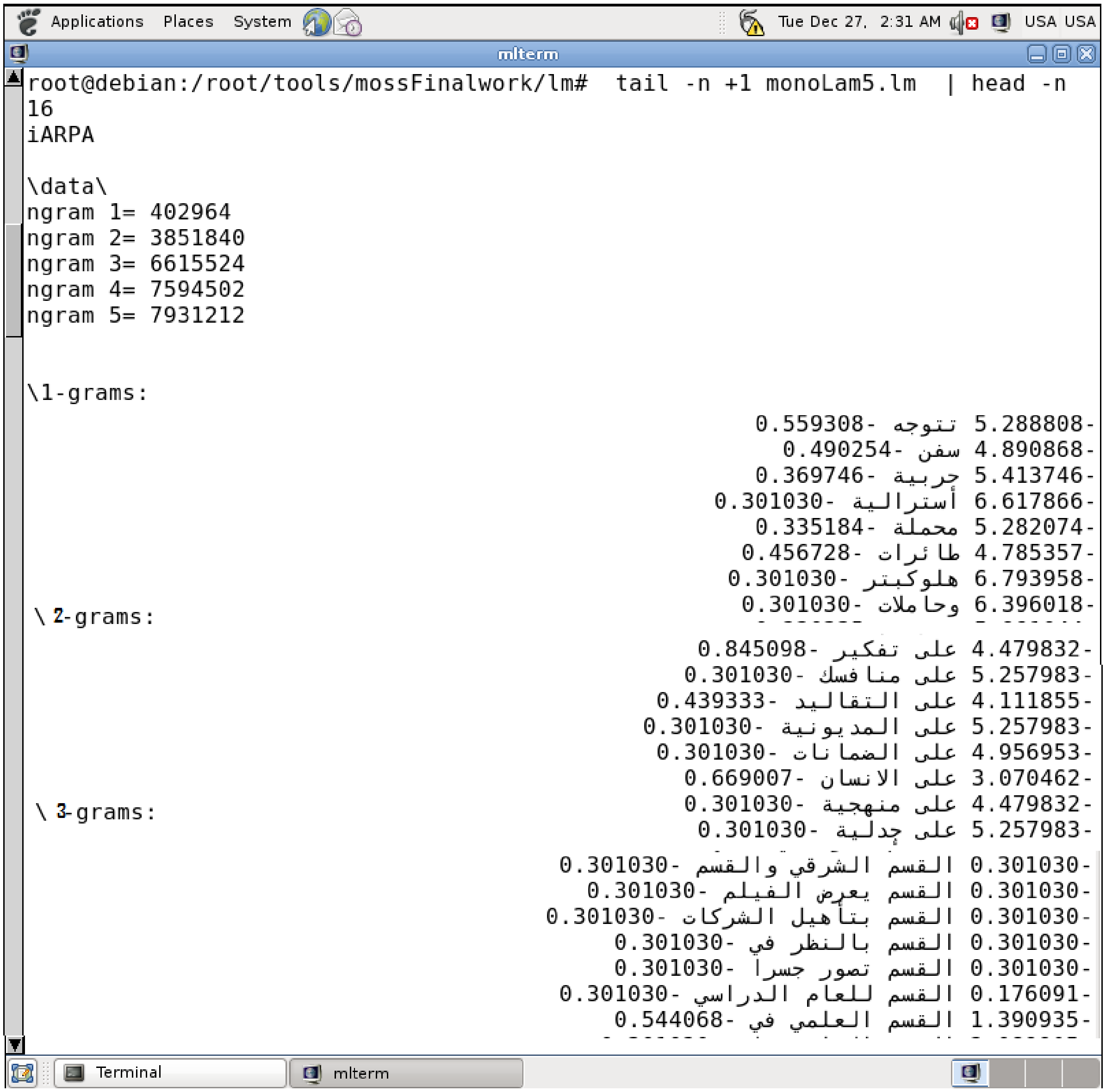


Figure 11: Language Model Development



```
root@debian:/root/tools/mossFinalwork/lm#  tail -n +1 monoLam5.lm  | head -n
16
iARPA

\data\
ngram 1= 402964
ngram 2= 3851840
ngram 3= 6615524
ngram 4= 7594502
ngram 5= 7931212


\1-grams:
-5.288808 تتوجه -0.559308
-4.890868 سفن -0.490254
-5.413746 حربية -0.369746
-6.617866 أسترالية -0.301030
-5.282074 محملة -0.335184
-4.785357 طائرات -0.456728
-6.793958 هلوكبتر -0.301030
-6.396018 وحاملات -0.301030
 \ 2-grams:
-4.479832 على تفكير -0.845098
-5.257983 على منافسك -0.301030
-4.111855 على التقاليد -0.439333
-5.257983 على المديونية -0.301030
-4.956953 على الضمانات -0.301030
-3.070462 على الانسان -0.669007
-4.479832 على منهجية -0.301030
-5.257983 على جدلية -0.301030
 \ 3-grams:
-0.301030 القسم الشرقي والقسم -0.301030
-0.301030 القسم يعرض الفيلم -0.301030
-0.301030 القسم بتأهيل الشركات -0.301030
-0.301030 القسم بالنظر في -0.301030
-0.301030 القسم تصور جسرا -0.301030
-0.176091 القسم للعام الدراسي -0.301030
-1.390935 القسم العلمي في -0.544068
```

Figure 12: Snapshot of the IRSTLM tool output of unigram, bigram and trigram.

4.4. Translation Model Development

Building the translation model requires the availability of Arabic to English parallel corpus, in order to generate the phrase table. Figure 13 outlines this process. In order to do so, we utilized an open source software called GIZA++ [42]. GIZA++ is an extension of the GIZA software that was developed by Fran Josef Och in 1999 at the Center of Language and Speech Processing of Johns Hopkins University [43]. This tool implements various word alignment techniques including the IBM models 1 through 5 and the HMM-based alignment model [42, 44]. Preparation of the training data involves storing the parallel corpus into two files: one for English sentences and the other for Arabic sentences. The files are sentence-aligned, i.e. each line in the English file contains one English sentence, with its corresponding Arabic translation sentence lying at the same line number in the Arabic file. Then, the text undergoes a data cleaning process similar to the one mentioned in Section B but at the word level instead of the character level. In addition, the training data is tokenized and all English text characters are converted into lower case characters. Finally, the phrase table is generated.

Table 5 shows the phrase table size generated by each of the three parallel corpora used in this thesis. The phrase table contains pairs of English phrases and their corresponding Arabic phrases, with the probability of each such pair, as show in Figure 14. The generated phrase table was directly used by the Moses decoder. However, it was not suitable for the Pharaoh decoder, and hence a script was written in order to make it suitable for the pharaoh decoder.

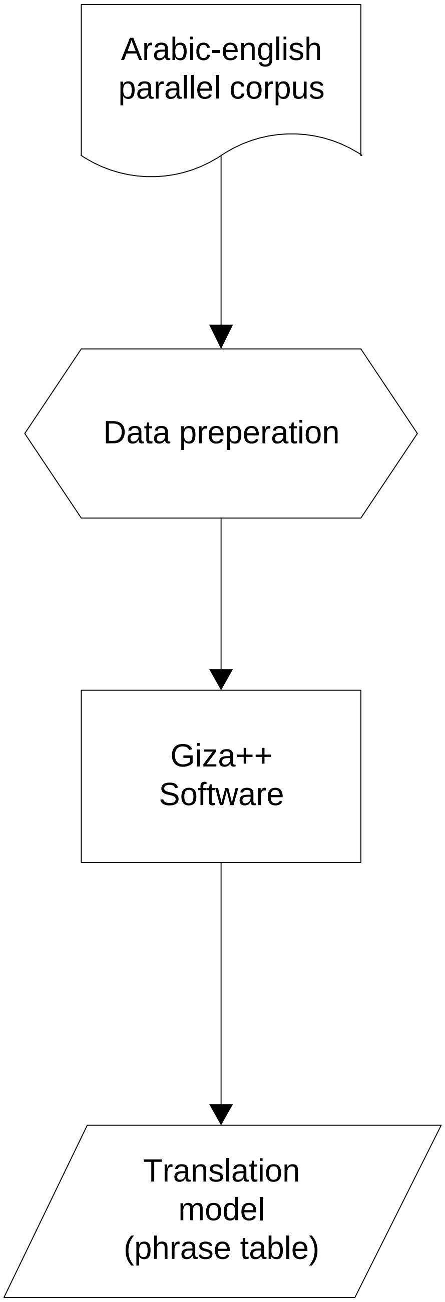
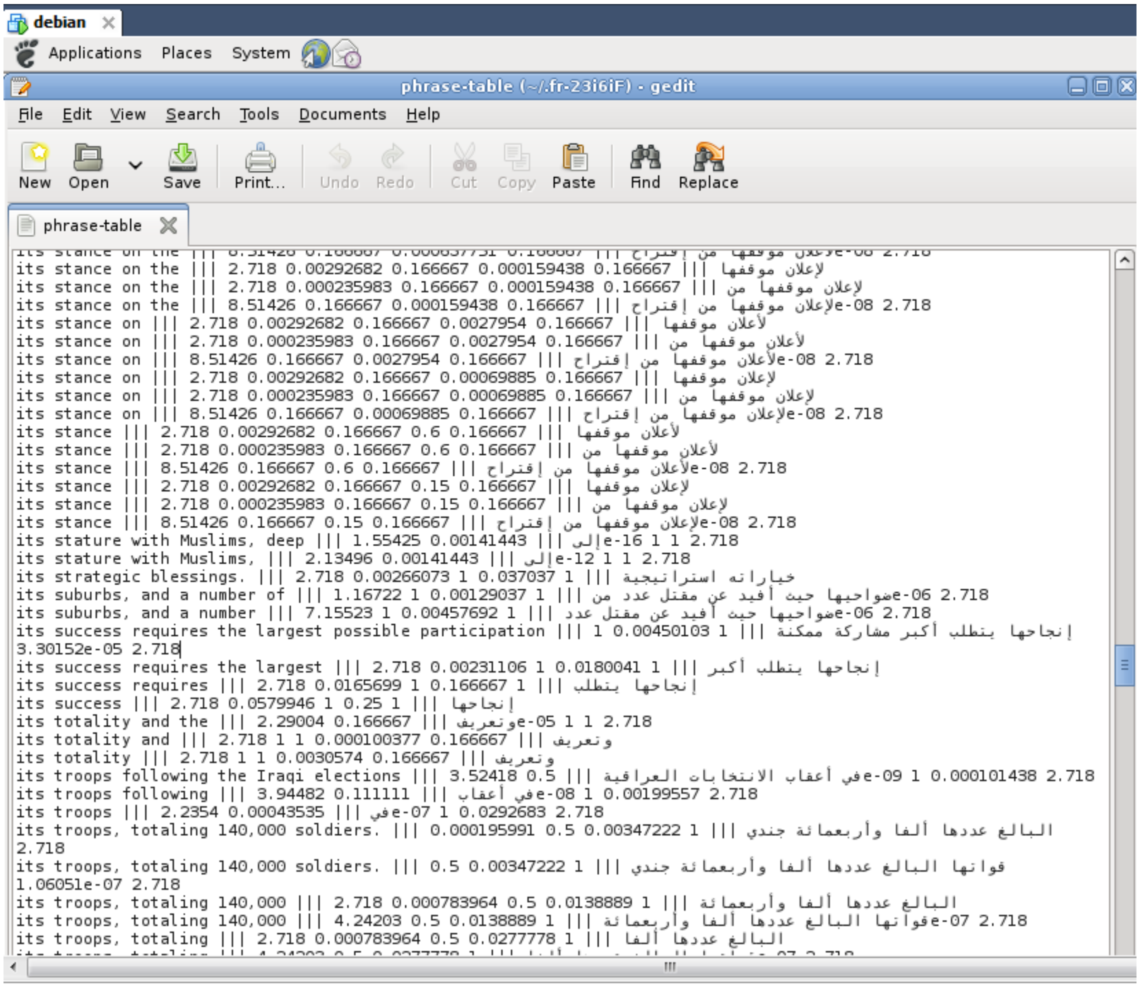

Arabic-english parallel corpus
Data preperation
Giza++ Software
Translation model (phrase table)

Figure 13: Translation Model Development.

Table 5: The corpora, their sizes and the sizes of their phrase tables.

Corpus	Corpus Size (in # of sentences)	Phrase Table Size (in MB)
LDC	3728	14.6
Meedan	17471	131.4
UN	45044	148.3



```
its stance on the ||| 8.51426 0.166667 0.000637731 0.166667 ||| لإعلان موقفها من إقتراحe-08 2.718
its stance on the ||| 2.718 0.00292682 0.166667 0.000159438 0.166667 ||| لإعلان موقفها
its stance on the ||| 2.718 0.000235983 0.166667 0.000159438 0.166667 ||| لإعلان موقفها من
its stance on the ||| 8.51426 0.166667 0.000159438 0.166667 ||| لإعلان موقفها من إقتراحe-08 2.718
its stance on ||| 2.718 0.00292682 0.166667 0.0027954 0.166667 ||| لأعلان موقفها
its stance on ||| 2.718 0.000235983 0.166667 0.0027954 0.166667 ||| لأعلان موقفها من
its stance on ||| 8.51426 0.166667 0.0027954 0.166667 ||| لأعلان موقفها من إقتراحe-08 2.718
its stance on ||| 2.718 0.00292682 0.166667 0.00069885 0.166667 ||| لإعلان موقفها
its stance on ||| 2.718 0.000235983 0.166667 0.00069885 0.166667 ||| لإعلان موقفها من
its stance on ||| 8.51426 0.166667 0.00069885 0.166667 ||| لإعلان موقفها من إقتراحe-08 2.718
its stance ||| 2.718 0.00292682 0.166667 0.6 0.166667 ||| لأعلان موقفها
its stance ||| 2.718 0.000235983 0.166667 0.6 0.166667 ||| لأعلان موقفها من
its stance ||| 8.51426 0.166667 0.6 0.166667 ||| لأعلان موقفها من إقتراحe-08 2.718
its stance ||| 2.718 0.00292682 0.166667 0.15 0.166667 ||| لإعلان موقفها
its stance ||| 2.718 0.000235983 0.166667 0.15 0.166667 ||| لإعلان موقفها من
its stance ||| 8.51426 0.166667 0.15 0.166667 ||| لإعلان موقفها من إقتراحe-08 2.718
its stature with Muslims, deep ||| 1.55425 0.00141443 ||| إلىe-16 1 1 2.718
its stature with Muslims, ||| 2.13496 0.00141443 ||| إلىe-12 1 1 2.718
its strategic blessings. ||| 2.718 0.00266073 1 0.037037 1 ||| خياراته استراتيجية
its suburbs, and a number of ||| 1.16722 1 0.00129037 1 ||| ضواحيها حيث أفيد عن مقتل عدد منe-06 2.718
its suburbs, and a number ||| 7.15523 1 0.00457692 1 ||| ضواحيها حيث أفيد عن مقتل عددe-06 2.718
its success requires the largest possible participation ||| 1 0.00450103 1 ||| إنجاحها يتطلب أكبر مشاركة ممكنة
3.30152e-05 2.718
its success requires the largest ||| 2.718 0.00231106 1 0.0180041 1 ||| إنجاحها يتطلب أكبر
its success requires ||| 2.718 0.0165699 1 0.166667 1 ||| إنجاحها يتطلب
its success ||| 2.718 0.0579946 1 0.25 1 ||| إنجاحها
its totality and the ||| 2.29004 0.166667 ||| وتعريفe-05 1 1 2.718
its totality and ||| 2.718 1 1 0.000100377 0.166667 ||| وتعريف
its totality ||| 2.718 1 1 0.0030574 0.166667 ||| وتعريف
its troops following the Iraqi elections ||| 3.52418 0.5 ||| في أعقاب الانتخابات العراقيةe-09 1 0.000101438 2.718
its troops following ||| 3.94482 0.111111 ||| في أعقابe-08 1 0.00199557 2.718
its troops ||| 2.2354 0.00043535 ||| فيe-07 1 0.0292683 2.718
its troops, totaling 140,000 soldiers. ||| 0.000195991 0.5 0.00347222 1 ||| البالغ عددها ألفا وأربعمائة جندي
2.718
its troops, totaling 140,000 soldiers. ||| 0.5 0.00347222 1 ||| قواتها البالغ عددها ألفا وأربعمائة جندي
1.06051e-07 2.718
its troops, totaling 140,000 ||| 2.718 0.000783964 0.5 0.0138889 1 ||| البالغ عددها ألفا وأربعمائة
its troops, totaling 140,000 ||| 4.24203 0.5 0.0138889 1 ||| قواتها البالغ عددها ألفا وأربعمائةe-07 2.718
its troops, totaling ||| 2.718 0.000783964 0.5 0.0277778 1 ||| البالغ عددها ألفا
its troops, totaling ||| 4.24203 0.5 0.0277778 1 ||| قواتها البالغ عددها ألفاe-07 2.718
```

Figure 14: English to Arabic Extracted phrase pairs .

4.5. Decoding stage

In the decoding stage, we find the best translation for the input sentence by examining all possible translations, giving a score to each one of them, and then selecting the highest score translation. This scoring depends on the language model and the translation model. This stage is the most critical stage, as there exists an exponential number of translations to choose from [18]. Among the widely used freely available decoding tools are the Moses decoder [43, 45] and the Pharaoh decoder[46]. The Moses decoder is an open-source toolkit licensed under the Open Source License for statistical machine translation that has been hosted and developed under sourceforge.net with an active research community [45]. The decoder was mainly developed by Hieu Hoang and Philipp Koehn at the University of Edinburgh and extended during a Johns Hopkins University Summer Workshop. It was further developed under EuroMatrix and GALE project funding [45]. The decoder was originally developed for the phrase model which implements a beam search algorithm. The Pharaoh decoder, on the other hand, is a closed-source tool that can only be used as a black box to carry out phrase-based MT. It was developed in the University of Southern California and the Information Sciences Institute as part of a PhD thesis by Philipp Koehn [46]. Both decoders were integrated into our system in order to compare and contrast the results produced by each one of them.

CHAPTER 5

Experimental Results

5.1 Introduction

In the previous chapter, we described in details the steps involved in developing a phrase-based SMT system. We also described the off-the-shelf tools used in this development. In this chapter, we elaborate more on the implementation details of our developed system followed by some experimentation in order to assess using the phrase-based SMT in translating English to Arabic.

With respect to the SMT system development, we first build three Arabic parallel corpora and one monolingual corpus. The parallel corpora underwent two data cleaning stages. Then, we developed the Arabic language model from the monolingual corpus. All software were compiled under Debian 5 Linux operating system. The Debian OS required an Arabic language package in order to be able to deal with Arabic text. Then, we were able to compile GIZA++, MOSES and IRSTLM and run them on our system. Since these tools have been developed by different researchers on different operating system versions, some problems were encountered during the compilation process. Upon further investigation, we discovered that some of these software tools require older versions of certain libraries, among them is the gcc compiler, which worked under the version 3.5.

5.2 Sample Test Set

In order to evaluate the performance of the developed MT systems, we generated 10 English sentences, where two sentences were chosen from each of the following 5 news categories: economic, technical, political, culture and sports. Figure 15 shows our test set. In order to run the system for the generated test set, the decoder need not consider phrases and/or words not appearing in the test set. Otherwise, the process will take extremely long time to execute (in terms of hours and days). For this purpose, a script was used to filter the phrase table. Table 6 shows the size of the filtered phrase table suitable for our generated test set. Finally, the decoder is executed on the filtered phrase table and the output translation is generated. Table 6shows that although the UN corpus is the largest, Meedan corpus generated the biggest filtered phrase table, indicating that it is more varied than LDC and UN corpus.

Table 6: Phrase table sizes for Moses and Pharaoh after filtering.

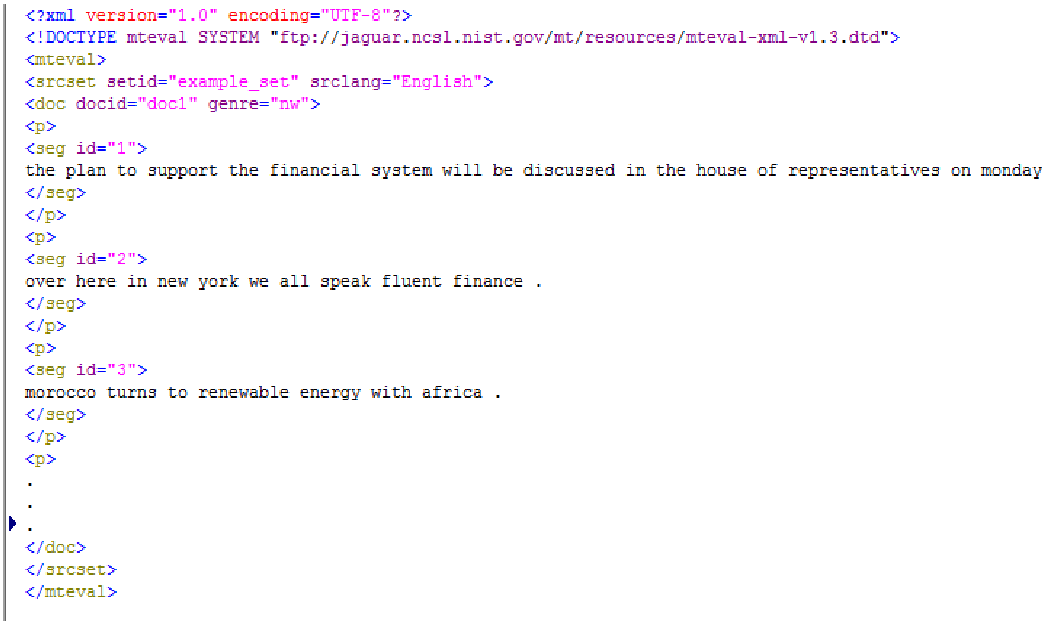
Corpus	Size of Phrase table	Size of filtered table for Moses	Size of filtered table for pharaoh
LDC	14.6 MB	264.1 KB	132.7 KB
Meedan	131.4 MB	4.8 MB	2.6 MB
Un	148.3 MB	2.3 MB	1.2 MB

English sentences	Categories
the plan to support the financial system will be discussed in the house of representatives on monday	Economic
over here in new york we all speak fluent finance .	Economic
morocco turns to renewable energy with africa .	Technical
this without any awful blue screen	Technical
international convention on the protection of the rights of all workers and their families .	Political
evolution instead of revolution .	Political
i want to be a young mother .	Culture
modern arab art museum opens in doha .	Culture
first class matches for the football starts today	Sports
viewers will be preoccupied with olympics until 2010	Sports

Figure 15 : English Sentences Test Set.

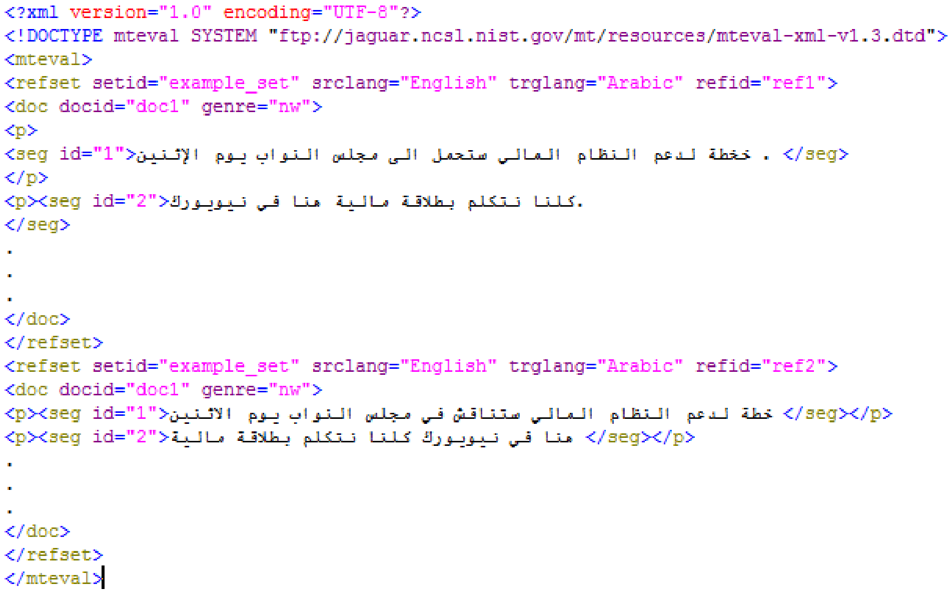
5.3 Performance Evaluation Technique

In order to assess the quality of the output translation, we need to employ an evaluation technique. The best evaluation approach is the human evaluation, which is very expensive and subjective evaluation. We utilized an inexpensive automatic evaluation approach called the Bilingual Evaluation Understudy (BLEU). This is a quick language-independent approach that correlates positively with the human evaluation. The BLEU score ranges from 0 to 1. Very few translations will get a score of 1, except when the translation is identical to a reference translation. Hence, a human translator may not reach a score of 1 [31, 47]. Therefore, we will evaluate our system using this approach and compare the results generated by Moses decoder and Pharaoh decoder. In addition, we will also compare the results generated by the three parallel corpora. we used a freely available script to calculate the BLEU score [48]. In order to run this script, we need three input files: the source file, the reference file, and the automatically generated translation file. The source file contains the original English sentences that need to be translated into Arabic. The reference file contains all possible translations of each sentence that were carried out manually. All these files must be written in XML form, as shown in Figure 16, Figure 17, and Figure 18.



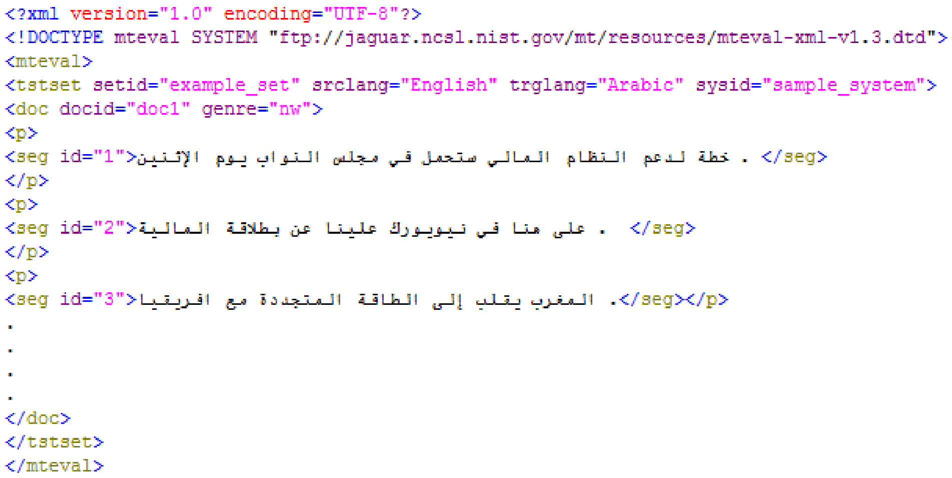
```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd">
<mteval>
<srcset setid="example_set" srclang="English">
<doc docid="doc1" genre="nw">
<p>
<seg id="1">
the plan to support the financial system will be discussed in the house of representatives on monday
</seg>
</p>
<p>
<seg id="2">
over here in new york we all speak fluent finance .
</seg>
</p>
<p>
<seg id="3">
morocco turns to renewable energy with africa .
</seg>
</p>
<p>
.
.
.
</doc>
</srcset>
</mteval>
```

Figure 16: Part of the source file that contains the original English sentences.



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd">
<mteval>
<refset setid="example_set" srclang="English" trglang="Arabic" refid="ref1">
<doc docid="doc1" genre="nw">
<p>
<seg id="1">خخطة لدعم النظام المالي ستحمل الى مجلس النواب يوم الإثنين . </seg>
</p>
<p><seg id="2">كلنا نتكلم بطلاقة مالية هنا في نيويورك.
</seg>
.
.
.
</doc>
</refset>
<refset setid="example_set" srclang="English" trglang="Arabic" refid="ref2">
<doc docid="doc1" genre="nw">
<p><seg id="1">خطة لدعم النظام المالي ستناقش في مجلس النواب يوم الاثنين </seg></p>
<p><seg id="2">هنا في نيويورك كلنا نتكلم بطلاقة مالية </seg></p>
.
.
.
</doc>
</refset>
</mteval>
```

Figure 17: Part of the reference file that contains all possible translations of the source file that were generated manually.



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE mteval SYSTEM "ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-xml-v1.3.dtd">
<mteval>
<tstset setid="example_set" srclang="English" trglang="Arabic" sysid="sample_system">
<doc docid="doc1" genre="nw">
<p>
<seg id="1">خطة لدعم النظام المالي ستحمل في مجلس النواب يوم الإثنين . </seg>
</p>
<p>
<seg id="2">على هنا في نيويورك علينا عن بطلاقة المالية .  </seg>
</p>
<p>
<seg id="3">المغرب يقلب إلى الطاقة المتجددة مع افريقيا.</seg></p>
.
.
.
.
</doc>
</tstset>
</mteval>
```

Figure 18: Part of the automatically generated machine translation output file.

5.4 Performance Evaluation Results

In our results we compared the evaluation of translation quality obtained by the SMT system using three different parallel corpora: LDC, Meedan, and UN. In the first set of experiments, we carried out the translation on Moses decoder and compared the results for the three corpora. Figure 19 shows the average BLEU score obtained for each news category for each corpus. It is clear that Meedan corpus scored best in economic, technical and culture news. In the political news, UN corpus scored the best. The LDC corpus scored a little better than Meedan in the sports category. With respect to the quality of the translation, it is obvious that political, economic and technical news scored reasonably well, whereas the translation was poor in the sports and culture categories. The results confirm that the Meedan corpus is the richest in content of different categories compared to the other two corpora. It also shows that specialized corpora, like that of UN, scored best in its domain, despite the reported filtered phrase table size is less than that of Meedan.

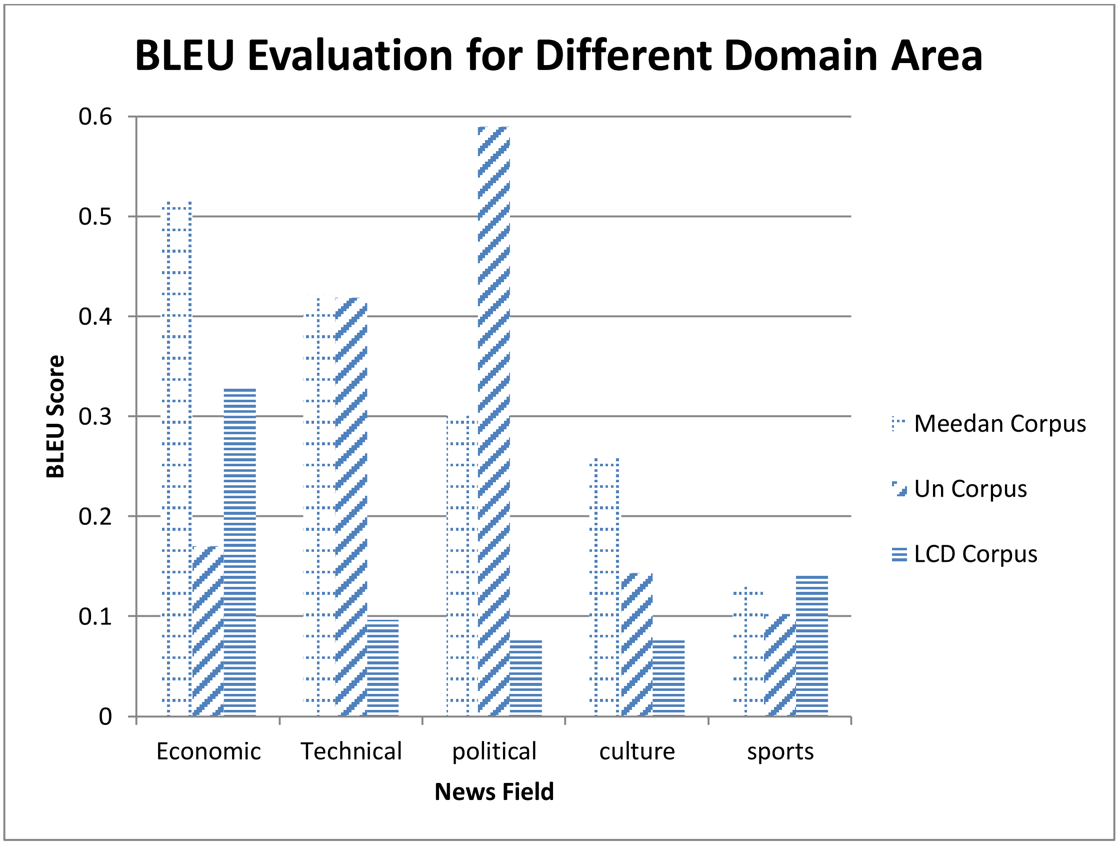


Figure 19: Evaluation Result of LDC, UN, Meedan Corpora for Different New Field.

Figure 20 shows the overall BLEU evaluation of the different corpora. It is clear that although the size of the UN corpus is bigger than that of the Meedan corpus, the Meedan corpus outperformed the other two corpora.

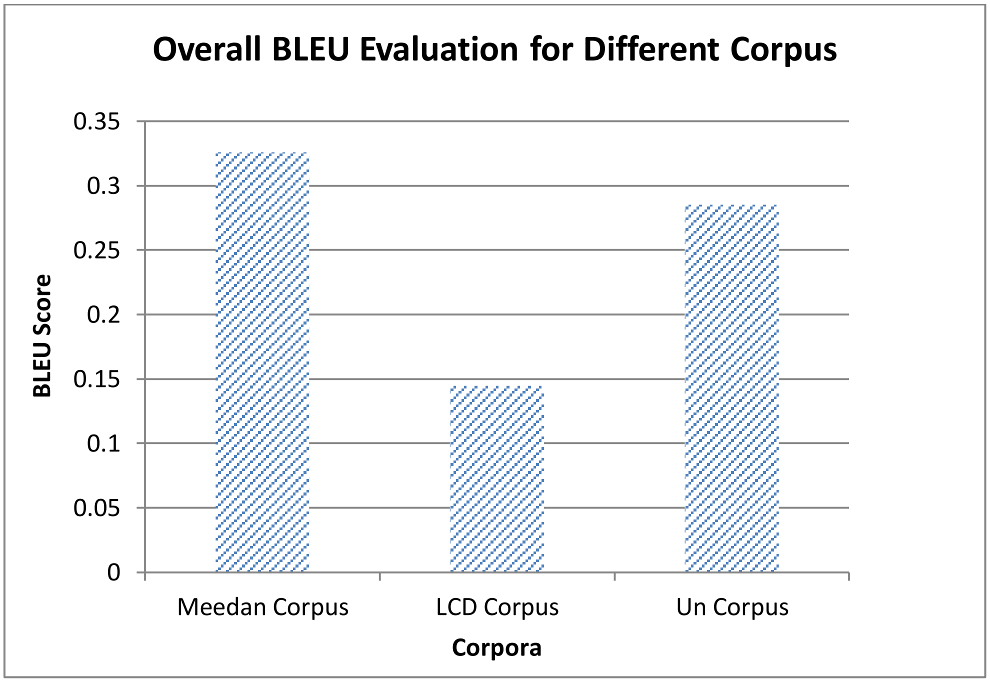


Figure 20 : Overall BLEU Evaluation for Different Corpus .

Next, we compare the performance evaluation of the SMT systems when run under different decoders, viz., Moses and Pharaoh decoders. Table 7 shows the BLEU score and the execution time needed to generate the results for the two decoders for all three corpora. The BLEU scores indicate that the Moses decoder produced better translations than Pharaoh decoder in all 3 corpora. The reason is that the Pharaoh decoder depends solely on the phrase translation probability P(*A*|*E*) to select the best phrases. In addition, it prefers Arabic translations containing the same number of words as that in the, corresponding source English sentence. The Moses decoder not only depends on the phrase translation probability P(*A*|*E*) to select the best phrase, but also uses additional information such as P(*E*|*A*) and the alignment information. As far as execution time is concerned, Moses performed slower than Pharaoh. Figure 21 shows that the Moses decoder achieved an overall BLEU score of 0.3261 on the Meedan corpus, which is greater than the 0.2152 BLEU achieved by the Pharaoh decoder.

Table 7: BLEU Evaluation And Execution Time For Pharaoh And Moses Decoders

	Pharaoh Decoder			Moses Decoder		
Corpus	LDC	MEEDAN	UN	LDC	MEEDAN	UN
Data Size	3728	17471	45044	3728	17471	45044
BLEU	0.11938	0.21521	0.21051	0.14453	0.32608	0.28503
Execution Time	16.74s	17.45s	18.53s	1m 45.69s	1m 58.72s	1m 50.38s

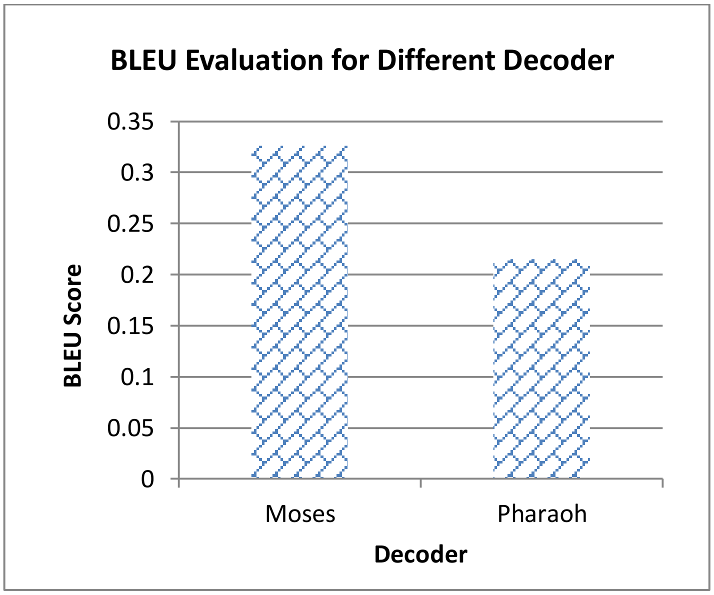


Figure 21 : BLEU Score for Moses and Pharaoh Decoder on Meedan Corpus.

Table 8 shown an example of the quality of the translation obtained from the different decoders based on the given reference translation.

Table 8: Example of translations obtained from different decoders

Original text	the plan to support the financial system will be discussed in the house of representatives on Monday
Reference	خطة لدعم النظام المالي ستحمل الى مجلس النواب يوم الإثنين .
Moses (Meedan corpus)	خطة لدعم النظام المالي ستحمل في مجلس النواب يوم الإثنين .
Pharaoh (Meedan corpus)	خطة لدعم النظام المالي ستحمل في مجلس النواب سانبيدج يوم الإثنين .

5.5 Limitations

In the carried out translations on the developed test set, we generally observed low quality sentence translations in all corpora. The reason for that is the problem of erroneous manual translation existing in the parallel corpora. For example, Figure 22 shows an English sentence and its corresponding translation from the Meedan corpus. It is obvious that the translation has been carried out in a summarization mode. Therefore, the facts that the attack was 80 miles off the cost of Gaza and the accusations of the activists were omitted from the Arabic manually generated translation. Hence, the existence of such translations degrade the quality of automatic translation. To overcome this issue, we have to remove all sentences that have been translated by summarization and possibly increase the size of the corpus.

The flotilla was attacked in international waters 80 miles off the Gaza coast, which activists say constitutes an act of piracy.
يذكر أن الهجوم حصل في المياه الدولية وليس في المياه الاقليمية لغزة أو اسرائيل.

Figure 22: Example of Bad Quality Translation in Meedan Corpus.

CHAPTER 6

Conclusions and Future Work

6.1 Summary and Conclusion

In this research work, an English to Arabic machine translation system was implemented based on the statistical phrased based approach. We gave an overview about machine translation approaches and state of the art research in machine translation (MT). The implemented system requires the development of a monolingual corpus that is used to generate the language model and a parallel corpus that is used to generate the translation model. Since the only available parallel corpora were either specialized, like the UN corpus or small, like the LDC corpus, we built a sizable parallel corpus spanning various categories of topics from the Meedan website, and later compared the results of Meedan with that of the other two corpora. The performance was compared based on the BLEU. Our experimentation shows that, overall, the Meedan corpus outperformed the other two corpora in most categories. We, also, compared the performance of the Moses decoder versus the Pharaoh decoder. We conclude that although the response time for the pharaoh decoder is better than that of the Moses decoder, the quality of the translation of the Moses decoder exceeds that of the Pharaoh decoder. Since there is a great lack in the availability of freely available tools and corpora that support research in the Arabic machine translation, we plan to make the monolingual and parallel corpora that we developed freely available over the web. We, also, plan to provide easy to use manual that outlines the steps needed to compile the various free tools used in this research to serve the Arabic language.

6.2 Future work

Our work in this research has opened many venues to explore and to improve on our achievements.

1) Part of speech tagging can be utilized in the decoding process in order to produce better translations. For example, an adjective follows the noun in the masculine or feminine type. Statistical translation techniques may fail in this regard.

2) In addition to statistical machine translation, one may use rule-based approaches that will improve the translation accuracy.

3) Using a large parallel corpus, that combines different domains will greatly improve the quality of the translation but may require high performance computing power.

4) Add new transliteration model for handling name translation. Such model will enhance the quality of name mapping from English to Arabic.

References

[1] A. Farghaly and K. Shaalan, "Arabic natural language processing: challenges and solutions," *ACM Transactions on Asian Language Information Processing (TALIP),* vol. 8, pp. 1-22, 2009.

[2] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, second ed.: Prentice Hall, 2009.

[3] M. I. Amro, H. Al-Muhtaseb, and W. G. Al-Khatib, "Development of Phrase-Based Statistical English to Arabic Machine Translation System Using a Parallel Corpus," *Journal of Communications and Computer Engineering,* vol. 2, p. 1: 5, 2011.

[4] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics,* vol. 19, pp. 263-311, 1993.

[5] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 2003, pp. 48-54.

[6] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics,* vol. 30, pp. 417-449, 2004.

[7] A. El-Desouki, A. A. Elgawwad, and M. Saleh, "A proposed algorithm for English-Arabic machine translation system," in *Proceeding of the 1st KFUPM Workshop on Information and Computer Sciences (WICS): Machine Translation, Dhahran, Saudi Arabic*, 1996.

[8] C. Pease and A. Boushaba, "Towards an automatic translation of medical terminology and texts into Arabic," *Proceedings of the Translation in the Arab World*, 1996.

[9] M. Ibrahim, "A fast and expert machine translation system involving Arabic language," Ph.D., Department of Applied Computing and Mathematics, Ph. D. Thesis, Cranfield Institute of Technology, UK, 1991.

[10] H. Mokhtar, Darwish, N., Rafea, A., "An automated system for English-Arabic translation of scientific texts (SEATS)," *International Conference on mMachine Translation and Multilingual Applications in the New Millennium,* November 20-22 2000.

[11] K. Shaalan, A. Rafea, A. A. Moneim, and H. Baraka, "Machine translation of English noun phrases into Arabic," *International Journal of Computer Processing of Oriental Languages,* vol. 17, pp. 121-134, 2004.

[12] K. Shaalan, A. Hendam, and A. Rafea, "An English-Arabic Bi-directional Machine Translation Tool in the Agriculture Domain," *Intelligent Information Processing V,* pp. 281-290, 2010.

[13] R. Al Dam and A. Guessoum, "Building a neural network-based English-to-Arabic transfer module from an unrestricted domain," in *Machine and Web Intelligence (ICMWI), 2010 International Conference on*, 2010, pp. 94-101.

[14] S. AWAD, *" فن الترجمة من الانجليزية الى العربية" (The Art of Translation between English and Arabic)*: Alraatib Aljame`yya, 1985.

[15] A. M. Mostafa, *"مرشدك الى الترجمة الصحيحة" (Your Guide to Correct Translation)*, 2000.

[16] E. A. Mohammed and M. J. Ab Aziz, "English to Arabic Machine Translation Based on Reordring Algorithm," *Journal of Computer Science,* vol. 7, pp. 120-128, 2011.

[17] M. F. Alawneh and T. M. Sembok, "Rule-Based and Example-Based Machine Translation from English to Arabic," in *Bio-Inspired Computing: Theories and Applications (BIC-TA), 2011 Sixth International Conference on*, 2011, pp. 343-347.

[18] P. Koehn, *Statistical machine translation.(English)*, 2010.

[19] A. E. K. N. Habash, "Orthographic and Morphological Processing for English-Arabic Statistical Machine Translation," *TALN 2010, Montréal*, 2010

[20] H. Al-Haj and A. Lavie, "The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation," *Machine Translation*, pp. 1-22, 2010.

[21] J. Elming and N. Habash, "Syntactic reordering for English-Arabic phrase-based machine translation," in *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, 2009, pp. 69-77.

[22] M. Diab, M. Ghoneim, and N. Habash, "Arabic diacritization in the context of statistical machine translation," in *Proceedings of MT-Summit*, 2007.

[23] H. Al-Muhtaseb. (2008). معالجة اللغة العربية والترجمة الآلية Available: http://faculty.kfupm.edu.sa/ics/muhtaseb/AC_Lectures_G081.htm

[24] Google. (2011). *Google Translate*. Available: http://translate.google.com/about/intl/en_ALL/

[25] Microsoft. (2012). *Microsoft Afkar*. Available: http://research.microsoft.com/en-us/projects/msafkar/

[26] Microsoft. (2011). *Bing Translator*. Available: http://afkar.microsoft.com/en/translation/

[27] ATA-Software. (2012). *Al-Wafi translator* Available: http://www.atasoft.com/documents/64.html

[28] R. Johnson and P. Whitelock, "Machine translation as an expert task," in *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Hamilton, New York, 1985.

[29] Meedan. (2010). *(Arabic → English) translation*. Available: http://news.meedan.net/index.php?page=events&post_id=300447

[30] K. Knight, "Decoding complexity in word-replacement translation models," *Comput. Linguist.*, vol. 25, pp. 607-615, 1999.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," presented at the Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania, 2002.

[32] A. Kilgarriff, "Web as corpus," in *Proceedings of Corpus Linguistics 2001*, 2001, pp. 342-344.

[33] D. Samy, A. M. Sandoval, J. M. Guirao, and E. Alfonseca, "Building a Parallel Multilingual Corpus (Arabic-Spanish-English)," in *Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC*, 2006.

[34] P. Resnik and N. A. Smith, "The Web as a parallel corpus," *Comput. Linguist.*, vol. 29, pp. 349-380, 2003.

[35] M. Abbas and K. Smaïli, "Comparison of topic identification methods for arabic language," in *International Conference on Recent Advances in Natural Language Processing-RANLP*, 2005.

[36] A. Rafalovitch and R. Dale, "United nations general assembly resolutions: A six-language parallel corpus," in *Proceedings of the MT Summit*, 2009, pp. 292-299.

[37] X. Ma, D. Zakhary, and L. D. Consortium, *LDC2007T24*: Linguistic Data Consortium, University of Pennsylvania, 2007.

[38] D. B. M. Abbas. (2010). *Khaleej-2004 ,Watan-2004 Corpora*.

[39] M. Madison. (2011). *The Web (Barely) Speaks Arabic*. Available: http://www.forumone.com/blogs/post/web-barely-speaks-arabic

[40] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Seventh International Conference on Spoken Language Processing*, 2002.

[41] M. Federico and N. Bertoldi. (2011). *IRST LM Toolkit*. Available: http://sourceforge.net/projects/irstlm/

[42] F. J. Och. (2000.). *GIZA++: Training of statistical translation models*. Available: http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

[43] F. J. Och and H. Ney, "Improved statistical alignment models," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, 2000, pp. 440-447.

[44] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, pp. 19-51, 2003.

[45] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, and R. Zens, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, 2007, pp. 177-180.

[46] P. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models," *Machine translation: From real users to research*, pp. 115-124, 2004.

[47] A. Agarwal and A. Lavie, "Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output," in *Proceedings of the Third Workshop on Statistical Machine Translation*, 2008, pp. 115-118.

[48] N. I. o. S. a. Technology. (2009). *BLEU scoring script for the Open MT Evaluation*.

Vitae

NAME: Mohammad Ismail Amro

BIRTH: 22th July 1984, Wadi-Alharia, Hebron, West-bank, Palestine.

NATIONALITY: Palestinian**, (**Jordanian Passports)

BACHELOR OF INFORMATION TECHNOLOGY

College of Administrative Sciences and Informatics, Palestine Polytechnic University, Hebron, West-bank ,Palestine.

MASTER OF SCIENCE IN COMPUTER SCIENCE

Department of Information and Computer Science**,** King Fahd University of Petroleum & Minerals, Dhahran, 31261, Saudi Arabia.

CONTACT: mohammad_amro@hotmail.com

MOBILE : +966535865614

ADDRESS: Masjid Uthman bin Affan,Wadi-Alharia, Hebron, West-bank, Palestine.