The Machine Translation of the Arabic tenses to the English tenses/aspects

Osama Zaki

MPhil.
University of Edinburgh
1999

Acknowledgements

I am grateful to Dr. Chris Mellish for his continuous support especially during my long period of study as a part time student. The work in this thesis would never have been completed without his valuable contribution all through my studies. I would like to thank Dr. Graeme Ritchie for his useful comments on my work and for his encouragement. Thanks also go to all the administrative staff at the department and to my colleagues.

Sincere thanks to my wife for her patience and for assisting me in writing readable English and correcting most of the early draft. I would like to thank my parents for their encouragement. Thanks to those who completed the questionnaire.

Finally, I would like to thank Dr Henry Thomson for his useful comments and valuable contribution and support after the exam.

Abstract

In many cases Arabic does not have any syntactic or morphological markers to signal the information that English conveys via the verb aspect (e.g. progressive or perfective forms). World knowledge and contextual knowledge are necessary to translate the Arabic tenses to the English tenses/aspects. In this work, a contrastive analysis between the two languages is carried out. Surface and non-surface linguistic information, in both languages, is studied and an attempt is made to examine the role of non-linguistic information. A module is implemented to automatically translate Arabic tenses into English tenses/aspects.

Declaration

I hereby declare that I composed this thesis entirely myself and that it describes my own research.

Osama Farouk Zaki
MPhil
Edinburgh
March 7, 1999

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Automatic translation from the Arabic tenses to the English tenses/aspects reveals interesting linguistic issues. We are particularly interested in the translation of two tenses: the Arabic Present tense to the English Present Simple and the English Present Progressive, and the translation of the Arabic Past tense to the English Present Perfect and the English Past Simple.

We believe that the automatic translation of these tenses represents a non-trivial and interesting problem for research since Arabic does not differentiate syntactically or morphologically between the Present Simple and the Present Progressive, and it also does not differentiate between the Present Perfect and the Past Simple. In fact the problem is the translation (mainly Arabic to English) of aspects rather than tenses because Arabic does not have explicit aspect markers.

Since the translation of these tenses requires non-linguistic information - world knowledge and context information - along with surface linguistic information, the raw output of an Arabic analyser is not sufficient for an English generator. Therefore a further process is needed for the Arabic analyser's output as the means of adding this extra information.

A module is implemented to automatically translate Arabic tenses into English tenses/Aspect. Neither an Arabic analyser nor an English generator is implemented. The assumption is made that the input to the proposed module is the output of an Arabic analyser and the output of the module, which is a decision about the English tense, will be used by an English Generator.

Initially, our application domain was a technical text, in particular text describing software, e.g. Windows 95, (e.g. sentences such as *The driver has detected an error*, *The sound card is not responding*, *An error occurred during installation*, *The field 'X' contains a character*). Nevertheless, because of the non-richness of the technical text (not enough variations in sentence structure) we had to consider text from religious sources.

1.1 Basic Linguistic Background

Arabic uses two main forms of the verb: the past form فَعَل *faʿala* (F1) and the present form يَفْعَل *yafʿal* (F2) to express all tenses. The past verb is used mainly in a situation that took place, or in a situation that was completed in an interval in the past. The present verb is used in a situation that currently takes/is taking place and will extend into the future [2] [6].

The Arabic past form and the present form have their equivalents in English, e.g. *did* and *do/does*. However, there are other forms in English which do not have an equivalent in Arabic, e.g. *doing* and *done*.

Unlike in English the two aspects, progressive and perfect, are not morphologically marked in Arabic (there are not special forms to express these as in English). A detailed linguistic analysis will be given within Chapter 2.

Table 1.1 shows that the Arabic Present tense can be translated to the English Present Simple, Present Progressive, Present Perfect and Present Progressive Perfect. The Ar-

Arabic	*English*
Present	Present Simple Present Progressive Present Perfect Present Progressive Perfect
Past	Past Simple Present Perfect Past Perfect
كَانَ *kāna* + Arabic Present	Past Progressive Past Perfect Progressive
Imperative	Imperative
سَوْفَ *swafa* + Arabic Present	will + Present Simple

Table 1.1: Tense mapping between English and Arabic

abic Past tense can be translated to the English Past Simple, Present Perfect and Past Perfect.

In Arabic the verb كَانَ *kāna* , sometimes, is used to express the progressive aspect in the past and the future but not in the present. In the present, verb semantics, context, world knowledge and/or gesture are used to express the present progressive. كَانَ *kāna* and/or the particle قَدْ *qad* are sometimes used to express the perfect aspect in Arabic [7] [9].

1.2 Translation Problems

Most of the Arabic sentences from (1) to (4) which contain the Arabic form (F2) and which refer to present time, syntactically, can be potentially translated to the English Present Simple or the Present Progressive.

(1) تستجيب بطَاقة الصوت للموجة

tstğyb bṭāqt alṣwt lllmwğt

respond card sound to-wave

The sound card is responding to the wave.

The sound card responds to the wave.

(2) يقرأُ برنَامج النسخ الاَحتيَاطي الملفَات

yqraʾu brnāmǧ alnsḫ alaḥtyāṭy almlfāt

read backup the-files

'Backup' is reading the files.

'Backup' reads the files.

(3) يحتوي الملف علَى بيَانَات معطوبة

yḥtwy almlf ʿlā byānāt mʿṭwbt

contain the-file of data bad

The file contains bad data.

The file is containing bad data*.

(4)يتطلب التطبيق إصدَارة 3.1 أَو مَابعدهَا

ʾaw mābʿdhā 3.1 *ytṭlb altṭbyq ʾiṣdārt*

require the-application version 3.1 or above

This application requires version 3.1 or above.

This application is requiring version 3.1 or above*.

Unless the translator has prior knowledge of the domain and the context in which the sentences appear as well as some other linguistic notions such as duration and perfection. any of these sentences will have two possible English translations.

The two English translations can never be equally good since the meanings of the two sentences to the reader of the English are distinct, while Arabic leaves the user to deduce from the available world knowledge the context of the sentence, or some other notions. However, each English translation not indicated by '*' would be acceptable in some context.

A similar problem, as shown in the following example, applies when translating the Arabic Past into English. The Arabic Past can be translated into either the Past Simple or the Present Perfect.

(5) حصلت مشكلة بَالجهَاز
ḥṣlt mšklt balǧhāz
occurred problem in hardware

A hardware problem occurred.
A hardware problem has occurred.

(6) كشفَ تفحّص الأَقرَاص عن تكوين غير صَالح
kšfa tfḥḥṣ ala'agrāṣ ʿn tkwyn ġyr ṣālḥ
detected scandisk configuration invalid

'Scandisk' detected an invalid configuration.
'Scandisk' has detected an invalid configuration.

(7) قمت بتثبيت برنَامج التشغيل
qmt btṯbyt brnamg āltṣil
done-you install driver

You have installed the driver.

You installed the driver.

The above examples show that the Arabic present tense can have at least two possible English translations and the same for the Arabic past tense. Non-linguistic information along with surface linguistic information is required to solve the translation problem.

It was also noticed that, by looking at the texts in our domain (detailed in Chapter 3), Arabic uses three particular verbs; يَجري *yağry*, يَقوم *yaqwm* and يَتم *yatm*, which we call Dummy Verbs to express time, and leaves the main verb in the infinitive form, as in (8). Another important issue is that sometimes what can be non-tensed (the infinitive) in Arabic can be tensed in English, as in (9). In our treatment of the translation problem, we will take into account these two phenomena.

(8) يَقوم تفحّص الأَقراص بكشف الأَخطَاء

yaqom tfḥ.ḥs alaʾaqrāṣ bkšf alaʾaḫṭāʾ.

‘yaqom’ Scandisk detecting the error

‘Scandisk’ is detecting the error.

’Scandisk’ detects the error.

(9) تَأَكِد من اتصَال جِهَازي الكمبيوتر قَبل نقل الملفَات

taʾaked mn itṣal ğehazy al-kmbytwtr qabl nql al-mlfāt

Make sure of connecting the two computers before transmitting the files.

Make sure you have connected the two computers before translating files.

Make sure you connected the two computers before translating files.

The non-tensed *connecting* can be translated into *have connected* and *connected.* But because of the phrase *make sure*, the first translation is the preferred one.

1.3 Review of Current Machine Translation Systems

Few attempts have been made at machine translation between Arabic and English and these have not been very successful. They are mainly commercial and are not based on theoretical studies.

At the time of writing this thesis, there are two machine translation systems (English to Arabic) on the market; ArabTrans and Al-Mutarjim Al-Arabey. There is also a machine translation system project under development at Al-Alamia. There is a general lack of material written on these systems.

Up till now and to the best of our knowledge, the only Arabic to English machine translation system is that by Aptek. Through a personal contact we have been able to obtain limited information about their treatment of the tense/aspect translation problem. They use Lexical-Functional Grammar (LFG) as their linguistic basis, to perform a multi-step transfer from the Arabic source to the English target. At one of these steps, the system generates an f-struc for the Arabic, then an f-struc for the target English. The f-struc consists of sets of featural pairs. The first element of the pair is the 'attribute' and the second element is the 'value'. Tense and aspect are features. For the sentence, *The manager had arrived from Rabat*, some of the featural pairs for the verb *arrived* would look as follows:

TENSE : past
PERFECTIVE : plus
PROGRESSIVE : minus

It is not clear, however, how they obtained these values from the Arabic verb وَصلَ *waṣla.* In other words what makes the value for the PERFECTIVE be 'plus' and not 'minus'? The English sentence could just as well be *The manager arrived from Rabat.*

Our work does not represent the development of a machine translation system, rather it is a module for solving the problem of the translation of tense/aspect between Arabic and English. It could be considered as an essential component of a machine translation system. Also, it could be a valuable contribution to an intelligent tutoring system for teaching English tense/aspect to Arabic speaking students.

1.4 The Domain and The Corpus

Technical text describing software can be found in various types. For Windows 95, from which our examples have been collected, there are four types: interface messages, help files, readme files, and the actual documentation.

The Help files in Windows 95 are in three styles: Procedural, Task, and Troubleshooter. Procedural files make many uses of the imperative (describing the steps to carry out a particular procedure, e.g. to set up a new device). Task files contain the text which appears in pop-up windows when a user enquires about a term and these texts are descriptive.

A random large mixture of these different styles of files were taken for the purpose of comparison and as a training corpora. However, as was mentioned earlier, further examples were taken from religious text to complement the weak formed technical text. The current training corpora comprises 107 examples.

The initial testing corpora was compiled from the troubleshooter files, since they have variations in sentence structure. Two files were chosen from the troubleshooter help files, amounting to a total of 6500 words. Then examples were taken from these two files (the Arabic text had been obtained by translation from the English by technical and professional translators). As for the training set, further examples from the same religious text were added to the testing corpora. The current testing corpora contains 18 examples.

Since the original text was written in English, the comparison between the Arabic and English is straightforward (i.e. no reverse translation has been carried out).

The examples were divided into groups according to their meanings (e.g. Conditional), which are outlined in Chapter 2. A manual comparison took place in an attempt to find regularities in the uses of tense/aspect in the two texts, English and Arabic, which would allow us to extract rules.

1.5 A Quick Overview of the Approach

The approach to the translation is based on applying rules to the information available about the Arabic sentence, i.e. the output of the Arabic analyser. These rules were extracted as a result of manual comparison of the same text in both languages.

A module was implemented. The module consists of the rules and the algorithm (the engine). The module takes as its input a representation of the output of an Arabic analyser with some other information added to it (such as the verb type of the English verb to be used and appropriate aspects of world knowledge), and it produces the proper English tense as its output.

1.6 Summary and Conclusion

This chapter represents an introduction to the project. We have looked at the scope of the work. We have shown how the translation of each of the two Arabic tenses, Present and Simple, can be translated into more than one English tense. The domain and the corpus were introduced. The computational approach was outlined. The stages of research can be summarised as follows:

1. Studying the verb, tense and aspect systems in both languages. This includes defining the terms tense and aspect.
2. Analysing the outcome of the research in contrastive analysis of the two tense systems.
3. Looking at the corpus, the training corpora, (a large non-specific number of files from our domain) and making comparison between both texts, then merging our results with the previous findings, step 1.
4. Suggesting a method for solving the problem of translating the tense/aspect. This is the essential part of the project.
5. Devising rules from the text based on the study which is carried out in 1, 2 and 3.
6. Implementing these rules and the rules interpreter.
7. Preparing another corpus for testing,*testing set.*
8. Testing and evaluating.

In conclusion, a separation between tense and aspect should be made at a semantic level in those two cases where there are neither morphological nor syntactic markers. A syntactic-based approach for translating the Arabic tense to the English tense/aspects is not sufficient on its own and it should be supported by a meaning-based approach.

In other words both surface linguistic information, such as syntactic information, and non-linguistic information are needed to translate the Arabic tenses into the English tenses/aspects.

In Chapter 2, we show how a meaning based approach can be reached by having a union of the tense meanings conveyed by both languages.

Chapter 3 describes how surface linguistic information can be exploited while Chapter 4 describes non-linguistic information. The computational approach is explained in Chapter 5. A summary and conclusion are given in Chapter 6. Finally further research and possible extensions are outlined in Chapter 7.