

Toward Corpus-Based Machine Translation for Standard Arabic

by *Mathieu Guidère, Ph.D.*

Abstract

The paper defines corpus-based machine translation and its possible applications in machine translation. The study is based on a bilingual corpus of French and Arabic texts and translation unit alignment. The criteria used for alignment combine linguistic and statistical information. The study also suggests procedures to build a machine translation system based on parallel translated corpora.

Key words

Corpus linguistics, parallel corpora, translated corpus, corpus alignment, machine translation, French-Arabic.

Introduction

The information revolution and technological innovations have driven the development of language industries and the expansion of multilingualism. The use of machine translation has experienced unprecedented growth with many diverse new techniques and demands. However, the prime objective of researchers and businessmen, in an Internet-dominated environment, has been the rapid development of translation systems that are both accurate and effective.

This technological development, along with the huge volume of translations available in different languages, point toward the use of this corpus for specific machine translation and computer-assisted translation applications.

The use of corpora of bilingual parallel texts seems to offer a promising tool for the future, thanks to the progress that has been made in terms of storage and computing capacities, as well as of acquisition of large amounts of text.

The idea of using parallel corpora is not new; it dates back to the early days of machine translation, but it was not used in practice until 1984 (Martin Kay 93). Subsequently, various methods have been proposed for processing the different levels of correspondence between two texts, an original and its translation.

The approach proposed here for the French-Arabic language pair (*corpus-based machine translation*) can be considered an extension of what was referred to, in the 1980s, as “memory-based machine translation” (MBMT) or “example-based machine translation” (EBMT)¹. It is based on a statistical approach making use of probability calculations of equivalences between texts of the corpus.

This method is grounded on the conviction that there are no preestablished solutions to translation (theoretical procedures), but most possible solutions can be found in texts already translated by professionals. In other words, a large portion of a translator's competence is encoded in the language equivalencies that can be found in already translated texts.

Moreover, a bilingual corpus is richer in information about the language than a monolingual corpus, since it provides situational equivalency information on the possibilities of the language system when in contact with a different linguistic system.

Different approaches to machine translation

A distinction can be made between two types of approach in machine processing of Arabic. On the one hand, approaches that can be qualified as “particularist” because they emphasize the linguistic idiosyncrasies of Arabic and use them for a local processing approach, which is considered more in agreement with the internal requirements of the Arabic linguistic system. On the other hand, the “universalist” approaches highlight the actual or assumed possibilities of application of methods already tested for other languages, such as English or French into Arabic, with or without adaptation.

This distinction is reflected, in each case, in research focused on specific points which make the two approaches different, although basically complementary. In fact, the “particularist” approaches are concerned mainly with the morphological and semantic aspects of the Arabic language, while the “universalist” approaches emphasize the syntactic aspects of the linguistic system. Although the two approaches are complementary, the distinction makes it difficult to coordinate research and use the results obtained by each one.

However, most research today focuses on a specific morphological aspect, the roots, and bases all considerations on the grammatical concept of “scheme,” which is peculiar to Semitic languages, and therefore not only to Arabic. The system thus obtained is a hybrid one: it involves the recursive application of a certain number of (basically morphological) rules stored in the machine’s memory. There is no true inference of interlinguistic representation, but, mainly, *no use is made of the corpus*. Nowhere is there a training process or integration of the translated text data that is not included in the rules of the system. How can a real-life text be expected to be handled in this case, since, by definition, the system is incapable of foreseeing-and solving-all the real translation problems occurring in the text?

Accordingly, we propose modeling the grammatical rules of Arabic to show how the “particularist” approach handles the problems of interlinguistic equivalence. In summary, the proposed system expresses the morphological and syntactic rules described by the traditional Arabic grammar books in computer terms. This research does not take into account the actual texts as they appear in translated bilingual corpora; it basically represents a model of the rules that govern the language, rather than research about translation and languages in contact. However, the advantage of such a system is that it provides a “rational” version of the phonological rules that are specific to the normative use of the language, which allows a comparison with the effective data of the corpus based on this initial research.

Actually, regardless of which approach is adopted, machine translation is never an objective *per se* or a priority. It is rather a latent aim of the work being carried out, but researchers prefer to concentrate on developing useful applications that will allow them to reach this goal eventually. In this research, machine translation is in fact a secondary concern, a situation that hinders the rapid development of effective systems.

There are currently very few applications for machine translation *into* and *from* Arabic², especially compared with other major languages such as English, French, or Spanish. The

few systems available primarily concern the Arabic-English pair and in reality constitute improved versions of electronic dictionaries³.

Other available applications (by Cimos and Alis, among other companies) have a restricted coverage of Arabic linguistic phenomena and are essentially based on specialized dictionaries. They are, in fact, technical translation aids rather than machine translation software packages.

In this context, the use of training corpora may advance machine translation research despite the fact that the corpora for Arabic currently present a few practical problems.

Bilingual corpora and associated problems

The parallel corpus consists of English or French texts, together with their translation into Arabic. In the current state of machine translation research *from* and *into Arabic*, no reference corpus is yet available . However, resources abound. Below we list a few examples of the available corpora:

- News agency wires (Reuter, BBC, AFP, etc.);
- Multilingual publications by international organizations (UN, Unesco, WHO, etc.);
- Foreign editions of biweekly and monthly periodicals (*Times*, *Le Monde Diplomatique*, *Elle*, *PC Magazine*, etc.);
- Classical and modern literary works translated from Arabic into English and French and vice-versa (for example, the catalog of specialized publishers such as *Maisonneuve* and *Actes Sud* in France).

It should nevertheless be noted that this considerable amount of text data has not been utilized so far because of certain practical problems, mainly the following:

1. *Acquisition*: Even when the corpora exist, they are not always available in electronic form or free of copyright.
2. *Conversion*: These corpora, of different origins and in various formats, must be processed to convert them into strictly linguistic corpora.
3. *Cleanup*: The corpora need human intervention to be put into a machine-useable format for processing.
4. *Synchronization*: The corpora must be aligned to identify the corresponding sections of the languages in question (English and Arabic or French and Arabic).

Not all of these operations are actually linguistic operations, but must be taken into account whenever a raw corpus is available.

We will focus here on the linguistic operations. They present various problems, which can be divided into three categories: ⁴

1. *Translation units*, i.e., the choice, definition, and delimitation of these units;
2. *Pairing*, i.e., alignment and synchronization of the translated corpora; and
3. *Algorithm*, i.e., the type of knowledge to be used in order to pair the languages (formal, lexical, semantic knowledge, etc.).

In the current state of research, the main difficulty for Arabic would appear to be category 1, i.e., defining the corresponding units of each corpus. Because once aligned, the two corpora will be submitted to a *document search*, the statistical module of which will reduce the secondary problems presented by the other levels of analysis.

Let us examine this question closer.

Alignment of a translated corpus

Aligning a corpus means making each translation unit of the source corpus correspond to an equivalent unit of the target corpus. In this case, the term “translation unit” covers both larger sequences such as chapters or paragraphs and shorter sequences such as sentences, syntagms or simply words.

The translation unit selected depends on the point of view chosen for the linguistic analysis and on the type of corpus used as a database. If the translated corpus demands a high level of faithfulness to the original, as is the case of mainly legal or technical corpora, the point of departure will be a close alignment of the two corpora, considering sentences, or even words, the basic unit. On the other hand, if the corpus is an adaptation, rather than literal translation of the original, an attempt will be made to align larger units such as paragraphs or even chapters.

The alignment operation can thus be refined based on the type of corpus. The linearity and faithfulness of the human translations help reliably align bilingual corpora. This is particularly true for predominantly technical corpora, but literary type corpora also lend themselves to reliable alignment of units below the sentence level if the types of equivalency observed on the corpus have been previously formalized.

It is obvious that the initial hypothesis, which allow these corpora to be used, is the correspondence-if not equivalence-both of the contents of the units considered and their mutual relationships. So-called “free” translations may present a serious processing problem: missing sequences, changes in word order, modification of content, etc. All these operations are quite common in everyday translation practice, but their frequency varies according to the field of the corpus.

Methodology

All these observations lead us to consider an aligned corpus not so much a set of equivalent sequences, but rather *corresponding text databases*. At any level (text, paragraph or sentence), the corpus should be considered a simple lexical database with “parallel units.” In other words, we propose a *search method* similar to the one used for *document information research systems* (with the help of a bilingual search engine).

Thus, the objective is not to show the structural equivalencies between the two languages, but, more modestly and more pragmatically, to search the T2 (target text) unit that is closest to the “request” which constitutes the T1 (source text) unit.⁵

To do so, the starting point may be a preliminary alignment of words with the help of a traditional bilingual dictionary (Deb 92). Such an alignment, although rough, may yield satisfactory results at the sentence level (Kay 93), especially when combined with a statistical method (Bro 93) and minimal formalization of the major syntactic phenomena.

The main advantage of this method is the use of translation memory, i.e., integration of the data which can be found in available text databases. The task can be somewhat simplified

by using specialized and regular fields of application as reference corpora. In fact, in specialized fields (legal, computer science, medical, etc.), the message will be “machine” translated essentially by using a customized basic dictionary and, above all, a translation memory created by the human translator during the training phase.

Another interesting aspect of the proposed method is the technique for searching the database. The application uses key words to retrieve the equivalent phrase segments in the two different texts/languages. Once they are found, they are formalized by a human translator as models before being stored in the translation memory. This type of procedure is recommended, but for the purpose of *automating the training process* (upstream) and not for validation (downstream). Here lies the major difference between machine translation (MT) and computer-assisted translation (CAT).

Linguistic analysis of the corpus

The different levels of linguistic analysis serve as a basis for the phase of automatic analysis and formalization of the translation equivalencies:

- First, morphological analysis identifies words or morphemes in the corpus.
- Second, syntactic analysis identifies syntagms and their functions.
- Third, semantic analysis identifies the meanings of the units and any ambiguities.

The morphosyntactic analyzer must ensure both accurate and effective analysis (quality and speed of processing). We therefore recommend a “superficial” syntactic description (*shallow parsing*), along with a “statistical” approach. In other words, the analyzer must preferably be supported by a grammar automatically acquired from the previously processed corpus.

The usefulness of such a corpus transcends its application for machine translation. While the main objective is create translation memories, other applications may also be considered, such as drawing up bilingual terminological lists and extracting examples for the purpose of computer-assisted teaching. Once annotated, the corpus could be used as the basis for the enhancement of electronic dictionaries, or to create grammar books.

In an ideal system, tagging should be performed automatically by comparing the texts of the corpus following a *probabilistic* procedure. Although sometimes an adjective can be translated by a noun or vice-versa, the categories given by reference dictionaries should help resolve such a situation. In this respect, the use of grammatical categories has a strong impact on the quality of tagging; a system with fewer categories seems to have a better success rate than a system with an exhaustive list of categories (Cha 95).

The empirical bilingual corpus approach

The probabilistic method will optimize the formalization of equivalencies in order to obtain the best possible machine translation.

The general idea of the procedure is to associate equivalent “translation units” (words, phrases, syntactic structures) with typical formal structures at the time the corpus sequences are identified and paired.

The basic purpose of such a procedure is to allow the pairing mechanism to be broken down into three parts:

1. Identify the *potentially* associatable units in the two corpora;
2. *Formalize* the structures of the associatable units using morphosyntactic tags;
3. Determine the *probability* of the proposed structures by comparing them to the effective data of the bilingual corpus.

By dividing the procedure into three phases, relatively simple translation models can be produced, so as to determine the units likely to correlate the theoretical analysis with the actual translations observed in the corpus.

One of the possible ways to make it easier to devise effective systems is to develop analysis methods based on the data stored in the *training corpus*. However, such methods, based on model training, depend on the amount of information available *a priori*, i.e., on syntactic rules previously developed by the human expert.

In this respect, a distinction can be made between two types of situations:

Situation 1: A parallel corpus of analyzed and annotated translation units is available *a priori*, i.e., a corpus for which a syntactic scheme representing the structure of the unit has been selected for each unit, given its meaning.

This first situation, where a considerable amount of information is available for estimating the parameters of the equivalency model, will be referred to as a *training* situation and will or will not be used depending on *how frequently it occurs* in the annotated corpus.

Situation 2: Only relatively scarce data, i.e., raw corpora, are available. In this case, equivalency hypotheses must be based on iterative re-estimations of the corpus data. For example, all units starting with the impersonal pronoun “one/on” may be grouped in order to compare their translations.

It should be emphasized that one of the advantages of the statistical model, compared to more theoretical approaches based on differential linguistics, is that it may considerably reduce the number of possibilities to be formalized. It is no longer necessary to consider all the rules of the language or all the possible translations of a sequence in order to process it, only apply *corpus-based equivalencies*.

Examination of a large sample of a bilingual French-Arabic bilingual corpus of the journalistic type available thus leads to the following interesting observations:

- First, for most of the corpus, a single unit of the target text corresponds to each source text unit.
- Second, the interrelationships between the units in the target text are almost the same as those in the source text, even if some sequences are sometimes inverted or omitted.
- Lastly, there are fixed reference points which mark the two texts and which allow rapid identification of the translation units. This is the case of numbers, dates, proper nouns, and titles, but also layout (for example, the division into paragraphs is often the same).

Based on the statistical analysis of the equivalencies at the word level using the document search method, the following types of equivalencies were identified:

- *Strong equivalence*: those rare cases where the number of words, their order, and their meaning in the (bilingual) dictionary are the same;
Example:
P1: “The increase in unemployment in the month of May is troubling officials.”
T1: “izdiyâd al-bitâla fî shahr mâris yuqliqu al-mas'ûlin.”
Literally: “the increase (in) unemployment in the month (of) May is troubling officials.”
- *Approximate equivalence*: cases where the number of words and their meanings are the same, but not the order in which they appear.
Example:
P1: “The President of the Republic received his Syrian counterpart.”
T2: “Istaqbala ra'îs al-jumhûriyya nazîrahu al-sûriyy.”
Literally: “received the President of the Republic his counterpart Syrian.”
- *Weak equivalences*: cases where the order and the number of words are different, but their dictionary meanings are the same.
Example:
P1: “Rains are being expected in the north of the country.”
T1: “Yutawaqqa'u an tumtira ghadan fî al-shamâl.”
Literally: “It is expected that it will rain tomorrow in the north.”

For the French-Arabic bilingual corpus available, most of the translation equivalencies were weak. Thus, the alignment of the corpus is based not only on the syntactic structure of the texts, but also on the semantic anchor points (pivot points). As long as two words in a sentence of the source text correspond to *at least*⁶ one word in a sentence of the target text, the two sentences are assumed to have an equivalency or translational relationship.

The reliability of such a search of the corpus is guaranteed, as noted above, by an intermediate alignment stage of the text at the paragraph or possibly at the sentence level. Thus, if two words appear in sentence S1 of text T1 and the search of the corpus makes two words of equivalent meanings appear in sentence S2 of text T2, the two sentence units are assumed to have a translational relationship.

To ensure the greatest possible reliability for the search operation, decreasing alignment of the bilingual corpus is used, from the largest translation units (chapters and paragraphs) to the smallest ones (sentences followed by syntagms and words). Thus, the field of analysis can be tightened by performing a “regressive” alignment of the corpus units and focusing the search on smaller and smaller units.

Corpus facts

In summary, the following hypotheses can be verified in most cases and on most of the corpus:

- Two chapters have a translational relationship if at least two paragraphs correspond to each other.
- Two paragraphs have a translational relationship if at least two sentences correspond to each other.
- Two sentences have a translational relationship if at least two words correspond to each other.

- Two words have a translational relationship if at least one of their meanings is confirmed by the bilingual dictionary used as a reference.
- It should also be emphasized that numbers (dates and figures) are fixed reference points in the two languages and reliable identifiers of the internal sentence units.
- There is little else but idiomatic expressions that pose a problem, especially when their position in the phrase is not the same in the two languages. However, using probability calculations, by a process of elimination, the equivalency of the remaining units can be confirmed after all the other units have been processed.

Concluding remarks

From a methodological point of view, combining a linguistic approach with a statistical approach makes it possible to fine-tune the alignment and enhance processing of bilingual corpora with a view to machine translation.

From a prospective point of view, the system requires only limited pretreatment (such as identifying and formalizing the “translation units”) and the use of a traditional bilingual dictionary (English-Arabic or French-Arabic). There is no need for long, exhaustive morphosyntactic tagging of each corpus. It is up to the machine to find the equivalencies by comparing the two corpora that have a translational relationship.

However, to ensure proper performance of the system, certain aspects warrant special mention.

First, the quality and size of the bilingual dictionary used must be considered from the outset. The dictionary may actually be very basic in terms of the grammatical information provided, but must be able to integrate the unknown words found in the bilingual corpus.

Second, the type of data used, i.e., the bilingual texts that are aligned, may pose a problem if the quality of the corpus is poor or if it has not been subjected to strict control by a human expert.

And third, the accuracy of the system and the quality of the translation depend on the volume of training data available and the accuracy of corpus synchronization.

For the above reasons, the first machine translation systems *from* and *into* Arabic cannot be expected to be infallible. A rather long training period on a large amount of different text data must be expected. Once this stage is completed, the information stored in the translation memory can be reactivated to yield all kinds of translation solutions that were previously the exclusive domain of human experts. However, to achieve this purpose, a dose of artificial intelligence will probably have to be integrated into a system thus designed.

Bibliography

The literature on research in machine translation *from* and *into* Arabic is scarce. The works cited below concern studies that may be used for developing specific applications.

Al-Daimi, K.J. & Abdel-Amir, M.A. (1994): “The Syntactic Analysis of Arabic by Machine”, *Computers-and-the-Humanities*, 28, 1, 29-37.

Brown, P. & alii (1990): “A Statistical Approach to Machine Translation,” *Computational Linguistics*, 16 (2) : 79-85.

- Brown, P. & alii (1993): "The Mathematics of Statistical Machine Translation : Parameter Estimation," *Computational Linguistics*, 19, 2.
- Catizone, R. & alii (1989): "Deriving Translation Data from Bilingual Texts," U. Zernick (ed.), Proc. of the First Lexical Acquisition Workshop, Detroit.
- Chanod, J.-P. ; Tapanainen, P. (1995): "Creating a tagset, lexicon and guesser for a French tagger," *Proc. of EACL SGDAT Workshop on Form Texts to Tags : Issues in Multilingual Languages Analysis*, Dublin, 58-64.
- Chen, K.H ; Chen, H.H. (1995): "Aligning Bilingual Corpora Especially for Language Pairs from Different Families," *Informations-Sciences-Applications*, 4, 2, Sept, 57-81.
- Darke, D. (1986): "Machine Translation for Arabic," *Language-Monthly*, 28, Jan, 10-11.
- Debili, F. & alii (1994): "De l'appariement des mots à la comparaison de phrases : un algorithme pour la reconnaissance de la paraphrase et de la traduction," *RFIA'94*, Paris.
- Debili, F. ; Sammouda, E. (1992): "Appariement des phrases de textes bilingues français-anglais et français-arabe," *Proc. 15th International Conference on Computational Linguistics (Coling 92)*, Nantes, France.
- Ibrahim S. & Mohammed M.A. (1993): "A Fast and Expert Machine Translation System Involving Arabic Language," *Dissertation-Abstracts*, 53, 9, Mar, 4777-B.
- Kaji, H. & alii (1992): "Learning Translation Templates from Bilingual Text," *Proc. 15th International Conference on Computational Linguistics (Coling 92)*, Nantes, France.
- Kay, M. & Röscheisen M. (1993): "Text-Translation Alignment," *Computational Linguistics*, 19, 1, Mar.
- Mankai, Ch. ; Mili Ali, (1995): "Machine translation from Arabic to English and French," *Informations-Sciences-Applications*, 3, 2, Mar, 91-109.
- Sato, S. ; Nagao, M. (1990): "Toward memory-based Translation," *Proc. 13th International Conference on Computational Linguistics (Coling 90)*, Helsinki, Finland.
- Simard, M. ; Foster, G. ; Isabelle, P. (1992): "Using Cognates to Align Sentences in Bilingual Corpora," *Proc. 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 92)*, Montreal, Canada.

¹ See Sato, S.; Nagao, M. (1990): "Toward Memory-Based Translation," *proc. 13th International Conference on Computational Linguistics (Coling 90)*, Helsinki, Finland.

² The most important companies working on machine translation applications into and from Arabic are Sakhr, Coltec, and Aptek.

³ Thus, the *al-Wafi* software by ATA, the software development company, works with the English language and is not based on true linguistic research (which shows in the result). Idem for the *SAT* and *CAT* systems by Sakhr or the *TranSphere* system by L&H Aptek.

⁴ See in this respect Debili F. (1997): “L'appariement: quels problèmes?” [Pairing: What Problems?], 1st JST Francil de l'Aupelf-Uref, Avignon, pp. 199-206.

⁵ See Hlal Y. and Alami Y. (1997): “Exploration de bases textuelles: emploi d'outils linguistiques” [Exploring Text Databases: Using Linguistic Tools] 1st JST Francil de l'Aupelf-Uref, Avignon, pp. 95-98, and Debili F. (1997): “Indexation interactive et interrogation multilingue fr-an-ar” [“Interactive Indexation and Multilingual Fr-En-Ar Searches”] 1st JST Francil de l'Aupelf-Uref, Avignon, pp.133-136.

⁶ This accuracy aims at integrating the Arabic models which translate form and semantic content.

Translated from French by Gabe Bokor
Edited by Alexandra Russell-Bitting