Morphological Analysis and Generation for Machine Translation from and to Arabic

Sadik Bessou
Ferhat Abbas University
Sétif- Algeria

Mohamed Touahria
Ferhat Abbas University
Sétif- Algeria

ABSTRACT
In this paper, we present machine translation importance and the need of a linguistic treatment for the transfer based approach, then we present our method in analysis and generation based on linguistic features of Arabic word, dealing with scheme concept; to extract morphological information, these information is very useful in tree generation and structural transfer.

Keywords
Machine translation, Morphological analysis, Morphological generation, Arabic.

1. INTRODUCTION

The importance of machine translation
Human translation cannot cope with the daily massive production of data. It cannot either deal with the technical material and terminology that are used consistently and need to be translated in the same way every day. Consequently, this leads us to envisage machine translation (MT) as probably the only reliable solution.

MT is not in itself an independent line of work; it draws from linguistics, computer science, artificial intelligence and translation theory ([13], [17]).

There is no doubt that Machine Translation has played an important role and will continue to evolve, given that we are living in a multi-cultural and multi-lingual environment. The existence of MT helps in bringing down the communication barriers.

With the dominance of English over other languages in many fields, it is obvious that more translators are required than before. The amount of translations carried out from English toward other languages is vast and is worth billions of dollars. The growth of the Internet and the computerization of the worldwide economy have changed the manner business is being conducted, emphasize the call for more efficient and faster techniques of translation, and manage to make use of the huge volume of available data online.

The need for a linguistic treatment before and after the transfer
If one wishes to have machine translation of high quality, there are two issues that must be resolved. The first of these, which must be solved before anything else, is the semantic aspect of the linguistic units. The second difficulty is that the translation calculus has to be adapted to the languages to be processed; very often machine translation needs special and appropriate calculi which suit the way a specific language works and subsequently the way the language can be processed according to another specific target language. [13]

2. MACHINE TRANSLATION
2.1 Machine Translation Approaches
Machine translation systems can be classified into three categories regarding their design: Direct, Transfer-based, and Interlingua-based systems. A general way to visualize these three approaches is to use the well-known "Vauquois Triangle" shown in Figure1 [18]. The triangle shows comparative depths of intermediary representation, with the Interlingua machine translation at the peak, followed by transfer-based, then direct translation. It presents the increasing depth of the required analysis on both the analysis and generation end as we move from the direct approach through the transfer approach, and to the Interlingua approach. Additionally, it gives one an idea about the decreasing amount of transfer knowledge required as we move up the triangle. From enormous amount of transfer at the direct level, where roughly for each word all knowledge is transfer knowledge, through the transfer level which needs transfer rules only for parse trees or thematic roles, and then to the interlingua level which symbolizes a language-independent conceptual structure of the source and target texts.[2]

The first approach, which has now been largely given up by MT researchers, consists of incorporating all the details for some specific pair of languages in one translation direction. Translation using the direct approach is done in almost a word-to-word manner. The second approach makes use of internal syntactic representations where knowledge is represented after disambiguation. The source text is first translated into an internal representation of the source language; this is then converted into an internal representation of the target language, which is finally used for the generation of text in the target language. The third approach makes use of an Interlingua, that is, an entirely independent language. The Interlingua is used as a common representation for the parsed source text as well as for the text to be generated in the target language. This approach, by far the most ambitious, has the advantage that it simplifies adding to a given MT system the capability of translating between additional pairs of languages. [19]
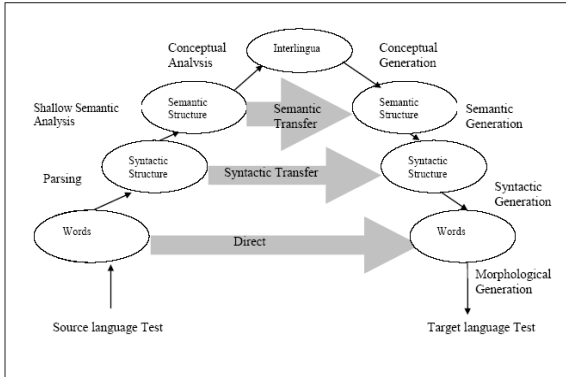
Figure 1. Vauquois Triangle: Machine Translation Approaches.
[18]

2.2 Transfer Based

In the Transfer approach, translation is completed through three stages: the first stage consists in converting source language texts into an intermediate representation, usually parse trees; the second stage converting these representations into equivalent ones in the target language; and the third one is the generation of the final target text [12]. In the transfer approach, the source text is analyzed into an abstract representation that still has many of the characteristics of the source, but not the target, language. This representation can range from purely syntactic to highly semantic. [2]

2.3 Linguistic knowledge

Many shortcomings in the output of MT are due to either faulty analysis of the source language text or faulty generation of the target language text. Enhancement to the output can be done only by formalizing our linguistic knowledge and enriching the computer with adequate rules to deal with the linguistic phenomenon. Fully automated, high quality machine translation (FAHQMT) has not yet been achieved. Yet there is a lot that we can do to improve the quality of MT output and increase its usefulness [1]

3. OUR METHOD

To avoid shortcomings in the output of MT we have to understand the text, so it is necessary to make a complete linguistic analysis. Linguistic techniques are theoretically the best because they provide a good understanding of the documents. However, implementation of such methods requires very powerful natural language processing algorithms, and a considerable processing time.

The suitable technique for the Arabic language is lemmatization; the stemming process converts documents to extract reference words. The role of stemming is to remove the last characters words (regarded as describing the words inflections). Some stemmings use more complete morphological knowledge (suffixes, prefixes…).

So instead of translating a word in its inflected form, we translate its lemma, then according to the features accompanying the lemma we generates the word in the target language.

The goal of linguistic analysis is to provide the words to translate into features that are used to generate the same type of matching words. The most important information is syntactic structures, which are driven from the different components of the word.

Our approach is to decompose words into morphemes [6], [16] taking into account the grammatical relationship between them. It proceeds first by a normalization that converts the document into a manipulable format [15]. This step is a delicate task because Arabic is inflectional and strongly derivational language [5].

The morphological analyzer can not function without the help of the dictionaries containing lexical units. This stage is lexical analysis to check if a lexical unit belongs to the language, but we must check the compatibility between the various constituents of the word.

After these two steps comes the third step that generate the lemma with all its features.

To verify that a word belongs to the Arab lexemes with the exception of proper names, some common names and tool words it suffices to find its corresponding root and scheme.

According to this method the proposed work will be based on three key steps:

Segmentation - schemes and roots research – lemmas and their features generation.

3.1 Segmentation

To resolve the ambiguity, Aljlayl and Frieder show that the light lemmatization (approach based on elimination of suffix and prefix) is significantly outperforms that based on detection of root in the field of information retrieval [3].

Cut out a word or segment it is to extract its various parts (Prefix, root, suffix …).

Segmentation principle is as follows:

- Segmenting the word in proclitic + base1+ enclitic which is to identify all proclitics and enclitics appearing in the word. Base1 (root + infixes) is usually fitted with a prefixes and suffixes.
- Segmenting base1 (result of the previous phase) in prefixes + base + suffixes; the principle of this phase is the same as the previous phase.
- Segmenting base in root and scheme i.e. find a scheme among the schemes stored in the scheme dictionary that matches the base. The method of scheme recognition will be described in 3.2.1.

3.1.1 Proclitics and Enclitics Recognition

Proclitics and enclitics research Fortunately the Arabic proclitics and enclitics list is limited. We can use the list proposed by [10] several have been used by [9] for lemmatization. The division of the word in " proclitics + base1+ enclitics" is not limited to proclitic research (enclitic respectively) in the word from the beginning to the end of the list, but to a certain compatibility between proclitics and enclitics identified in the segmented word.

Compatibility Test After the extraction of the proclitic P and the enclitic E from the analyzed word, these two substrings are merged into a string C to test them in a list of incompatibilities. If the C string is not found then this proclitic P is compatible with this enclitic E. [7].

Analysis principle When segmenting the word in: proclitic+ base1 +enclitic the process identifies the longest proclitic (enclitic respectively) in the word, and then accesses the table to check the compatibility between the two (proclitic and enclitic). If consistent the decomposition is accepted, it will

be stored in the table of results of this phase, then continues with a new decomposition to treat all possible cases. Otherwise the decomposition is wrong, we move to another one.

3.1.2 Prefixes and Suffixes Recognition The principle of this step is practically the same as the previous except that the compatibility table used is the ones of prefixes and suffixes.

3.2 Scheme and Root Research

3.2.1 Scheme Research For a word X, a scheme I correspond to this word X if the scheme length is equal to the word length and if all the letters corresponding to the positions in the field (list of infixes) in the scheme dictionary are found in the word X in the same positions revealed by the field. (figure2). [8].
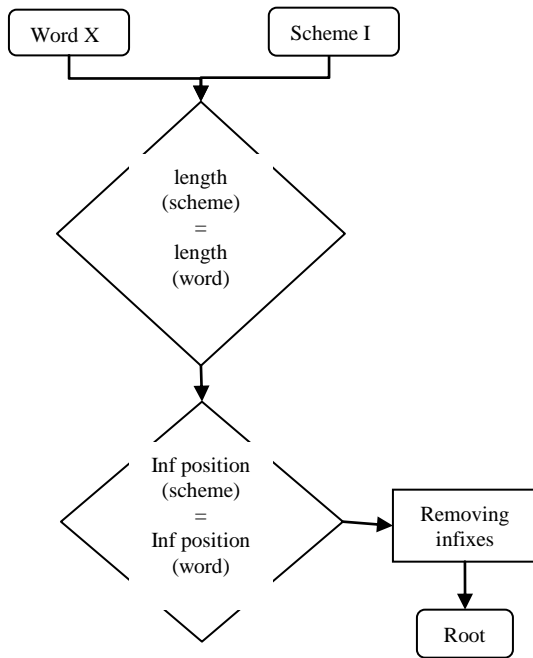

Figure 2. Scheme and Root Research.

Example Here is an example that helps us to understand the process:

Word= 'صالح', "salih", the process of finding scheme consults all the records that have the same length as the word to find the appropriate scheme: 'فاعل', "fail" The corresponding infixes list field is '2', the letter 'ا' is at the second position of the word 'صالح' so it is probably the right scheme.

3.2.2. Root Research After determining the scheme, the root extraction is limited to the removal of all the letters corresponding to the positions of infixes list field in the analyzed word.

Example The word 'مفاتيح', "mafatih", has as scheme 'مفاعيل', "mafail", the infixes list field is '135'. (figure3).

The suppression of the letter 'م' from the position 1 of the word 'مفاتيح' that it is the same as the infixes list field, the letter 'ا' from the position 3 and the letter 'ي' from the position 5 give

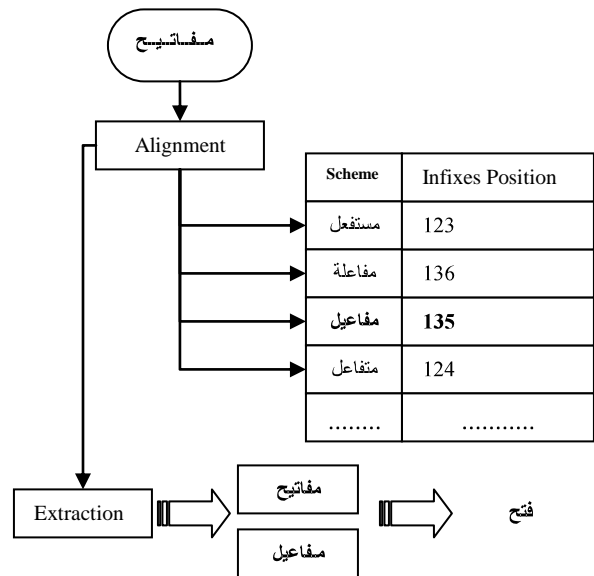'فتح',"fataha". Thus we found the correct root of the word 'مفاتيح' which is the root 'فتح'.

Scheme	Infixes Position
مستفعل	123
مفاعلة	136
مفاعيل	**135**
متفاعل	124
....	....

Figure 3. Scheme and root research and root extraction.

Remarque These steps are not the only ones, and sub-steps may be introduced if necessary (compatibility processing of proclitic / prefixes and enclitics / suffixes, desambiguïsation where there are several interpretations).

3.3 Features Calculation

The recognized proclitics and enclitics list as parts of analyzed words without ambiguity in the final analysis and after compatibility tests are stored, the same thing regarding prefixes and suffixes. This information is very useful to represent the word features and preserve the morphological characteristics of the word.

Example The word (أستخرجانها), "assatokhrijaniha", is in English (Are you going to take it out). The word is analyzed as lemma and other components that are interpreted as features that will be used in translation in the target language (figure4).
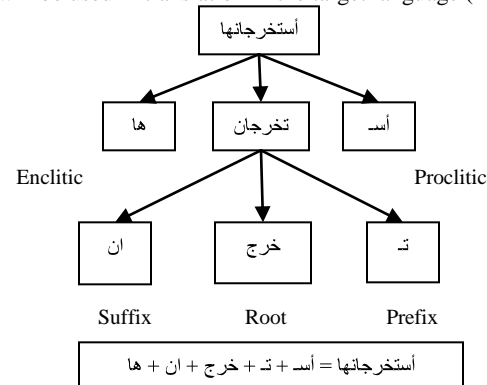

أستخرجانها = أسـ + تـ + خرج + ان + ها

Figure 4. Morphological Analysis of the Word 'أستخرجانها'

The various components have the following features:

Table 1. The Word Component Features.

Component	Feature
(أ)	Interrogative pronoun
(س)	Tense = future
(ت)	Subject = 2ⁿᵈ person singular
(ان)	Subject = dual
(ها)	object = feminine singular

These features are used in translation as follows:

Table 2. Using Features

feature	Target word
Interrogative pronoun	Are you
Tense = future	going
Verb = see dictionary	Take out
Subject = 2ⁿᵈ person singular	You
Subject = dual→plural	You
object = feminine singular	She, it

So we have the sentence (Are you going to take out it), then, using syntactic rules of agreement and word order, the sentence will be rewritten according to the grammar rules of the target language.

Remarque The feature of number can be applied automatically, but not gender, because the last has relation with cultural aspect of each language.

3.4 Generation

The generation of words in Arabic is used in cases where the target language is Arabic. The principle is the same but this time the process is done in the contrary, i.e. from a source language word with its features we generate the word in Arabic. First we translate the root in the Arabic language, and then we apply transformations to the root through the concept of scheme, finally according to features we add prefixes and suffixes. (figure5).
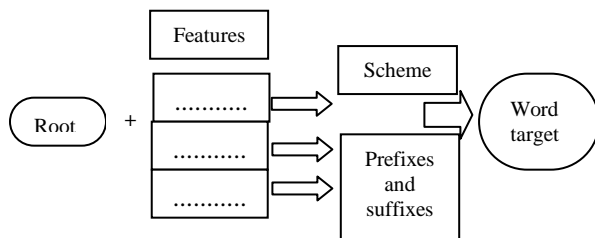

Figure 5. Word Generation from Root and Features.

Example The word (the workers) has the following features : noun, definite, masculine, plural, so the equivalent of the root

(work) is (عمل) after the application of features on this root (agent, doer we apply the scheme « فاعل », definite we add « ال », masculine we do not add « ة », plural we add « ون ») we obtain the word (العاملون).

4. CONCLUSION

We have presented our method based on scheme concept in the Arabic morphology; this method gives benefits both to analysis and generation for machine translation process. The syntactic Information is useful in generating internal representation to proceed with structural transfer.

5. REFERENCES

[1] Abu Shquier. M, Sembok. T, Word Agreement and Ordering in English-Arabic Machine Translation, IEEE 2008.

[2] AlDam Rasha Samih, Building a Transfer Module from Restricted Domain English-to-Arabic Bilingual Corpus, Master Thesis, University of Sharjah, January 2008.

[3] Aljlayl, M. and Ophir, F. On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach. In 11th International Conference on Information and Knowledge Management (CIKM), November 2002,pp. 340-347.

[4] Alsharaf H, Cardey S, Greenfield P, Shen Y, Problems and Solutions in Machine Translation Involving Arabic, Chinese and French, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), IEEE 2004.

[5] Attia, A. A large-scale computational processor of the Arabic morphology. A Master's Thesis, Cairo University, (Egypt), 2000.

[6] Attia, A. Devloping Robust Arabic Morphological Transducer Using Finite State Technology. In 8 th annual CLUK Research Colloquium, 2005.

[7] Bessou, S., Louail, M., Refoufi, A., Kedem, Z., Touahria, M. Un système de lemmatisation pour les applications de TALN. CITALA 2007. Rabat, Maroc.18-19 juin 2007. pp.35-51.

[8] Bessou S, Saadi A, Touahria M, Vers une recherche d'information plus intelligente application à la langue arabe, in the proceedings of the first international conference on information systems and economic intelligence, SIIE2008 , Hammamet, Tunisie, 14-16 Février 2008, pp. 91-100.

[9] Chen, A. and Gey, F. Building an Arabic Stemmer for Information Retrieval. Proceedings of the Eleventh Text Retrieval Conference (TREC 2002). National Institute of Standards and Technology,2002, pp. 631-640.

[10] Darwish, K. Probabilistic Methods for Searching OCR-Degraded Arabic Text, Doctoral dissertation, University of Maryland, 2003.

[11] Jisha P.Jayan, Rajeev R R, S Rajendran, Morphological Analyzer and Morphological Generator for Malayalam – Tamil Machine Translation, International Journal of Computer Applications(IJCA),2011.

[12] Hutchins J., Machine Translation: A Brief History, Concise History of the Language Sciences: From the Sumerians to the Cognitivists. Koerner E. F. K. and Asher R. E. (ed.). Oxford: Pergamon Press, pp. 431- 445, 1995.

[13] Hutchins J., Machine Translation and Computer–Based Translation Tools: What is Available and How it is Used. In: Bravo, J.M. (ed.): A New Spectrum of Translation Studies, University of Valladolid, Spain, 2003.

[14] Hutchins W. J. and Somers H. L., An Introduction to Machine Translation, London: Academic Press, pp. 56-57, 1992.

[15] Larkey L S, Ballesteros L. and Connell M. , Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland, 2002, pp. 275-282.

[16] Mohamadi, T., Mokhnache, S. Design and development of Arabic speech synthesis. WSEAS 2002, Greece, September 25-28, 2002.

[17] Senellart J., Dienes P., and Varadi T., New Generation Systran Translation System. MT Summit VIII Santiago de Compostela Spain, September 2001.

[18] Vauquois B., La traduction automatique à Grenoble. Paris: Dunod, 1975.

[19] Zantout R., and Guessoum A., Arabic Machine Translation: A Strategic Choice for the Arab World, journal of King Saud University, Volume 12, 2000, pp. 299-335.