Arabic to French Machine Translation System based on DCF Approach

Fahima Bouzit
Computer Science Department
Badji Mokhtar University
Annaba, Algeria
Bouzit.fahima@gmail.com

Mohamed Tayeb Laskri
Computer Science Department
Badji Mokhtar University
Annaba, Algeria
laskri@univ-annaba.org

***Abstract*— Machine translation is one of the disciplines that cover the automatic language processing domain. There are two different approaches to realize a machine translation system; linguistic approach, based on a set of theories and rules that govern the processed language and statistical one, based on probabilities and mathematic theories. Our system emerges from the linguistic school and contains in fact tree modules, each one concern a step in the natural language process; lexicon-morphological, syntactic and, semantic analyze(which we would describe in this paper).**

Keywords- Natural language processing; machine translation; linguistic approach; semantic analysis;

I. INTRODUCTION

Machine translation is the translation by a machine of a text written in a natural language (source language) to another language (target language) [1]. It is one of the disciplines that cover the automatic language processing domain. In this paper, we will describe our idea which aims to develop a system for Arabic-French machine translation using the DCF approach (which combines several methods that emerge from language school). We will show the usefulness of linguistic theories including: the method of conceptual dependencies, the theory of Fillmore, the semantic features of Shafe and frame based representation of Minskey, and how they were combined to achieve a translation taking into account, the context and meaning.

II. FILLMORE THEORY

Verbs differ according to their topological character, for example, certain verbs necessitate the semantic cases: 'Agent' and 'Object' where others need other cases such as 'Source' and 'Destination'. The cases ideally form a unique, finite, small in number, universal and valid list in every language [2]. For the Arabic language, semantic cases are drawn using casual marks, for example, the case 'Agent' is recognized by the grammatical case: Nominative, marked with the diacritic '‘' or the suffixes '‌ان ـ' or '‌ون ـ' . And the case 'Instrument' is recognized by the grammatical case: Dative, marked with the diacritic ' ـِ' or the suffixes '‌ـِين' or '‌ـِين ' , and preceded by the preposition ' بـِ 'or the words 'بواسطة', 'باستعمال' ,etc.

The advantage of this method is that it permits to make a representation of the sentence that does not stop in the syntactic analyze results limits [3], in other words, even if two sentences have different representations, they may transport the same sense. For example, the sentences:

- أمن المهندس فاعدة المعلومات
- أمِنت قاعدة المعلومات من طرف المهندس

The subject is different in despite the action (verb) is the same.

المهندس and فاعدة المعلومات play the same syntactic role: subject, where the agent is in the two cases المهندس and the object is always: فاعدة المعلومات.

III. SEMANTIC TRAITS OF SHAFE

This method consists to endow every noun in the definitions dictionary with many semantic traits showing relations it can have with the other words used with it in the sentence.

For noun representation, Chafe proposed a classification model. He defined a list of semantic traits (markers) that represent noun proprieties. According to Chafe, the noun is characterized with the traits: Animated, Human, Feminine, Unique, Concrete, Countable and Potent. [4] and the traits Consumable and dimension could be added [5].

Examples:

المستعمل=[(+)Animated,(+)Human,(-)Feminine,(-)Unique, (+)Concrete,(+)Countable,(+)Potent,(-)Consumable,(-) Dimension]

الشاشة=[(-)Animated,(-)Human,(+)Feminine,(-)Unique, (+)Concrete,(+)Countable,(-)Potent,(-)Consumable,(-) Dimension]

Although this method was conceived and used just for nouns,we proposed that it can be applied on verbs to resolve the problem of information lack we meet if the user wants to translate a non-vowelized text (we have to note that short vowels play a very important role in the Arabic language disambiguation). Consequently if the user wants to translate the sentence أرسل طفل رسالة إلكترونية إلى الأستاذ, the system can recognize the agent which is طفل thanks to the semantic traits of the verb أرسل [+ human], which means that this action can be down only by a human agent and so, the sy stem verifies

the traits of every noun in the sentence: رسالة [-human], طفل [+human]. Consequently, the agent can't be other than طفل.

IV. Frame Based Representation

Once the semantic cases drawn, we must find a way to represent them and the relations between them. A multitude of choices are available, but given the characteristics of the Arabic language, (you can swap the components of the sentence without affecting the meaning), we chose the method proposed by Minskey: frames. Frames have a whole set of slots reserved for the different concepts may contain a sentence, what drives us therefore to provide a slot for each component that may be encountered.

In general, each verb has its own characteristics and therefore requires a reduce number of slots [5]. and we know that there are verbs that require the same slots and that most verbs are used to express ideas that may well be expressed by other basic verbs. This leads us to use a verb classification method, we have chosen to use the theory of conceptual dependency.

V. Conceptual Dependency

This theory is characterized by the axioms:
1. Two sentences having the same meaning in one language or two different languages (although they have very different syntactic structures), should have the same internal representation. ;
2. Any information implied in a sentence must be made explicit in the representation

3. Any action is expressed in terms of primitives. Each primitive have an associated schema which must be instantiated and filled (at least partially) during the understanding process.

For example the primitive of the verb drink is the same one for the verbs eat or swallow (INGEST primitive) [4]. This theory therefore allows the reduction of each set of verbs in a primitive which shall be the representative and will now undergo a common treatment for all these verbs instead of duplicating it for each one. ,
In our work, we considered the eleven basic actions proposed by Schank.

So for two verbs that refer to two similar actions, we use the same primitive, for example, two verbs that denote an action of transfer of something abstract (eg possession), verbs such as: أخذ, سلم , أرسل , ATRANS primitive is used. Thus, we have the same frame that represents these verbs. The difference lies in the contents of the Action field. We can implement the frame as a list, table, or an implementation more efficient and organized using Objects. In our system, a class was defined for each primitive and during the frame construction phase, we instantiates an object of the class to which the word belongs and fills its fields.

VI. Required Informations

To have all the information about the different words which carries the analysis, it is necessary to have a part in the dictionary (a field or table) that contains all the information that can help make a good salary, To do this, we ranked the words in the dictionary in four tables: Verbs, Nouns, Adjectives and Particles. For example, the table contains the tense of verbs and its basic primitive but also the various semantic features, and its translation.
But during the implementation, it was noted that there are special cases that must be treated separately. Example: let the words:

لوحة = panneau ; تحكم = contrôle ; مفاتيح = clés.

The verbatim translation (from Arabic to Ftench) of لوحة التحكم and المفاتيح لوحة gives: لوحة التحكم = panneau de configuration and لوحة المفاتيح = panneau de clés. While the translation of لوحة المفاتيح should be: clavier.

The solution proposed was to put these strings of words in a table called Sequences and see during processing (step construction of the frame) if the text contains one of these suites which case we put directly its translation (the equivalent word or sequence in the target language) in the target frame. For example the words: ماسح ضوئي = scanner, سلة المحذوفات = corbeille,...

And therefore the number of tables used by the analysis module and the module for word translation is five: Nouns, Verbs, Articles, Adjectives and Sequences.

VII. Translation Running Description

When we enter a phrase to translate, the system analyzes it. In fact, the analysis goes through three phases: a morpho-lexical analysis that aims to recognize each word in the sentence, a parser to pull the various syntactic cases (subject, object, COD, COI,...). The results of this analysis are the inputs of the next phase: the semantic analysis as described in tfig1.

The system recognizes the action after it compares the words of the sentence to translate with the entries of the table of verbs, then consults with the primary field (class of verb) to extract the type of verb (Atrans, PTrans, ...), and the system starts filling the fields after instantiation of the class of the primitive.

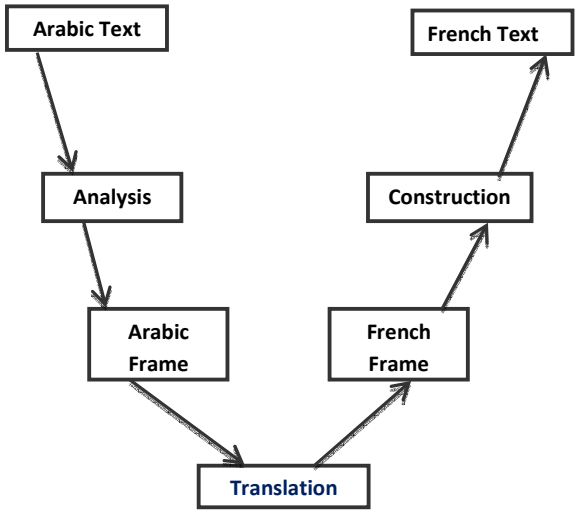

Figure 1. System global architecture

At this stage, the definition dictionary is consulted to extract the semantic traits of each word of the sentence in order to analyze it:

• The subject is recognized for example by the casual mark 'Damma', and according to the semantic traits of verb and different noun groups of the sentence. /

• The instrument is recognized by the particles or words: (بـ , بواسطة,…). .

• The source and destination with particles: (من) and (إلى , نحو, صوب). .

• Concerning the adjectives and particles, they can be recognized easily (a table is devoted to the adjectives and an other one to the particles), etc..

After the frame of the Arabic sentence (Arabic frame) is created, comes the role of the word by word translation module, so we get the destination frame (French frame). Then, the system dials the sentence in French from the target frame in an order that had been previously defined: Sentence = Agent Action Object Source Destination Beneficiary Instrument...

The system provides the result after the organization of the sentence according to the rules of syntax and grammar of the target language (taking into account the type and number of nouns and verb tenses: Présent / passé / futur), the gender and number of nouns,...

All these operations (generation of the sentence from the target frame and its organization to meet compliance of the target language) are provided by the module of management of the French language.

We note that the translation produced by the linguistic system (based on the DCF approach) goes through three basic phases: analysis, word for word translation and generation.

Some examples of translations made by the system are given in the following figure, fig2

VIII. CONCLUSION

Our translation system that some modules were exposed in this article, takes part in the semantic processing of texts using purely linguistic tools and finds fulfillment with the DCF method as a basis.

This method has been proved highly adaptable to the Arabic language and its particularities as to syntax and semantic sides [1],[3],[5],[6].

We can underline as a perspective for this work, to integrate to the system a good morphological analyzer (such as the open source tool Aramorph) and enrich the dictionaries used to cover other application areas and improve the results, because most dictionaries are richer and more defined rules are detailed, the resulting translation will be more accurate.

Figure 2. Example of translation

REFERENCES

[1] S. Russel, et Al. Intelligence Artificielle avec prés de 400 exercices, Pearson Edition, 2005.

[2] K. Meftouh, K.Smaili, M.T. Laskri, “Extraction automatique du sens d’une phrase en langue Française par une approche neuronale,” JADT, 2002.

[3] R. Mahdjoubi, “système pour le traitement automatique de la langue arabe basé sur ces propres caractéristiques,”Mémoire de magister, 1994.

[4] M.T. Laskri, “Sémantique du langage naturel à travers un système support de thésaurus,” These de doctorat d’état, 1994.

[5] K.Meftouh, “Un réseau simplement récurrent pour la génération d’une représentation du sens d’une phrase écrite en langue arabe basée sur les cas sémantiques,” Mémoire de magister, 2000.

[6] K. Rezeg, “Une Approche connexionniste pour la traduction automatique des textes arabe en français,” Courrier du Savoir. N°08, pp.59-67, 2007.