ATRANF: MACHINE TRANSLATION SYSTEM PROTOTYPE
(APPLICATION ON ARABIC TO FRENCH TRANSLATION)

Fahima Bouzit, Mohamed Tayeb Laskri Badji
Mokhtar University, Algeria
Bouzit.fahima@gmail.com, laskri@univ-annaba.org

ABSTRACT
This work falls within the framework of natural language processing. Our goal is to develop a machine translation system from Arabic to French using a purely linguistic method as a key to get insights into the standard layer-based structure of linguistic phenomena (morphology, syntax and semantics) as well as in recognizing the interaction between them, which we think is the most appropriate to a such rich in morphology and syntax language as Arabic. To ensure that, we used a set of linguistic theories such as: Fillmore theory, conceptual dependency, semantic traits of Chafe and frame representation of Minsky. We will show in this paper, the usefulness of these methods and how we combined them to realize a multilingual translation system.

Keywords: machine translation, Arabic, semantic cases, Fillmore theory.

1 INTRODUCTION

Machine translation has for decades attracted the interest of researchers in artificial intelligence, which gave rise to two different currents. The current based on linguistic theories, where the principle is to identify all the rules and features of the language, in order to use them to create dictionaries and rules, and use a set of formalisms to move from the external representation of the sentence, to an internal universal one. The second current consists to use large amounts of aligned bilingual text to estimate the probabilities of models; this is the statistical or probabilistic current.

In this paper, we will describe our idea which aims to develop Arabic to French machine translation system using a purely linguistic approach, which combines several methods as Fillmore Theory, the semantic features of Chafe and the frame based representation of Minsky.

In the linguistic approach, one of the most important phases is the semantic analysis, which involves extracting the meaning of surface structures using a variety of tools and methods.

To understand the meaning of a sentence, it is essential to know the meaning of its various components and the role of each one of them [1]. In order to ensure this, we used Fillmore theory. The idea is to consider the verb as the kernel of the sentence and to study the role of its other constituents (nouns) with this kernel.

2 FILLMORE THEORY

Verbs differ according to their typological characteristics, for example, there are verbs that require the semantic cases: 'Agent' and 'Subject', although some other verbs require other cases, such as 'Source' and 'Destination'. The cases ideally form a single, limited, small in number, universal and valid list in all languages [2]. For the Arabic language, semantic cases are identified by casual marks (short vowels), for example, the case agent is marked by the nominative grammatical case marked by the diacritic 'ُ' or the suffixes 'ان' or 'ون'. Where the case instrument is recognized by the dative case 'ِ' or the suffixes 'ُّ' or 'ًُّ' and is preceded by the preposition 'ـ' or the words 'ل', 'ا', etc.

The advantage of this method is that it allows to give to the sentence, a representation that does not stop at the tips of the results of syntax parsing [3], in other words, even if two sentences have different representations, they may transport the same meaning. For example, the sentences:

- ذا إ و أر (The child sent an e-mail to the teacher).
- ار ذ ١ (an e-mail was sent to the teacher by the child).

The subject is different, although, the action (verb) is the same, and the words و (e-mail) and ١ (child) play the same syntactic role: subject, while the agent is in both cases ١ (child) and the object is always: ذ (e-mail).

We extracted and specified the Arabic semantic cases, based on its characteristics [1], here are some examples:

- The case AGENT:
  Syntactic Case = Subject.
  (grammatical case=Nominative)

UbiCC Journal Page 26 www.ubicc.org
The case OBJECT:
Syntactic Case = Object Comp
Or
Syntactic Case = Subject
Verb mode = Passive

The case INSTRUMENT:
Grammatical Case = Dative
Preposition = ا،ـ،ل

The case SOURCE:
Grammatical Case = Dative
Preposition = ْ◌ِ◌،إ،َ◌،َ◌ْب
Or
A place noun playing the role of a direct object complement of some known verbs, such us: در،ك،ـ،ى

The case DESTINATION:
Grammatical Case = Dative
Preposition = ـِ◌،إ،َ◌،َ◌ْب
Or
A place noun playing the role of a direct object complement of some known verbs, such us: در،ك،ـ،ى

3 SEMANTIC TRAITS OF CHAFE

This method consists to endow every noun in the definitions dictionary, with many semantic traits, showing the relations it may have with the other words used with it in the sentence [4].

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]:

• لـ [{(+) Animated, (+) Human, (-) Feminine, (-) Unique, (+) Concrete, (+) Countable, (+) Potent, (-) Consumable, (-) Dimension}]

• ..: ..: [{(-) Animated, (-) Human, (+) Feminine, (-) Unique, (+) Concrete, (+) Countable, (-) Potent, (-) Consumable, (-) Dimension}]

Although this method has been developed and used only for names, we proposed to apply it on verbs to solve the problem of the lack of information that occurs if the user wants to translate a text without short vowels. So if the user wants to translate the sentence: لـ ظـ ءـ ؛، the system can identify the agent which is “لـ ظـ ءـ ؛” through the semantic features of the verb ظـ ءـ ؛ [+ human], which means that this action ظـ ءـ ؛ can be made only by a human, and therefore, the system checks the features of every noun in the sentence: ظـ ءـ ؛ [-human], لـ [human], and consequently decides that, the agent can only be لـ ظـ ءـ ؛.

4 FRAME REPRESENTATION

Once the semantic cases drawn, we must find a way to represent them and the relations existing between them. There is a wild choice of methods and formalisms for this (Context Free Grammar, Recursive Transition Networks, logic grammar, knowledge based processing,…), but given the characteristics of the Arabic language: we can swap the components of the sentence without changing its meaning. For example: the six sentences are correct, and transport the same meaning:

- لـ ظـ ءـ ؛
- ظـ ءـ ؛ لـ
- ظـ ءـ ؛ لـ
- ظـ ءـ ؛ لـ
- ظـ ءـ ؛ لـ
- ظـ ءـ ؛ لـ

So we can’t use any of those formalisms, because we will be face to huge and difficult to build representations and grammars. This leads us to choose the method proposed by Minsky: frames [1].

Frames have a whole set of slots, reserved for the various concepts contained in the sentence to represent, what drives us to provide a slot for each component that may be encountered Fig. 1.

![Figure 1: Representation, in the basic frame, of the sentence “لـ ظـ ءـ ؛”](image-url)

One can easily notice that there are several slots that are empty, and this is in fact, the drawback of this type of representations: the waste of storage space, because in general, each verb has its own characteristics and therefore requires a reduced number of slots [1].
We know that there are verbs that require the same slots as others, and that most verbs are used to express ideas that may well be expressed by other basic verbs. This leads us to use a method of classification of verbs. For that reason, we chose the theory of the conceptual dependency. So, the sentence $\text{the child easily printed the text with the printer}$, for instance, will be represented in a frame where the number of slot is smaller (a specialized frame), Fig. 2.

**Table 1:** the eleven basic primitives proposed by Shank.

<table>
<thead>
<tr>
<th>Primitive</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROPEL</td>
<td>Apply a force to something</td>
</tr>
<tr>
<td>MOVE</td>
<td>Moving a body part</td>
</tr>
<tr>
<td>GRASP</td>
<td>Catch an object Ingest, for a moving object</td>
</tr>
<tr>
<td>INGEST</td>
<td>Physically expel, for a moving object</td>
</tr>
<tr>
<td>EXPEL</td>
<td>Modify an abstract relationship, such as possession</td>
</tr>
<tr>
<td>PTRANS</td>
<td>Produce a sound, support of an action such as &quot;communicate&quot;</td>
</tr>
<tr>
<td>ATRANS</td>
<td>Information transfer</td>
</tr>
<tr>
<td>ATTEND</td>
<td>Apply his attention to a perception or stimulus</td>
</tr>
<tr>
<td>MTRANS</td>
<td>Creating a new though</td>
</tr>
<tr>
<td>MBUILD</td>
<td></td>
</tr>
</tbody>
</table>

So for two verbs referring two similar actions, we use the same primitive, for example, for any verb that denotes an action of transfer of something abstract (eg possession), such as the verbs: $\text{MTRANS}$, $\text{ATRANS}$, we use the primitive $\text{ATRANS}$. 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 using objects. We defined a class for each primitive, and during the frame construction phase, we instantiate an object of the class to which the verb belongs and fill its fields.

**6 DICTIONARIES**

To have all the information about the different words needed during the analysis, it is necessary to have in the dictionary (fields or tables) that contains any information that could be useful.

To insure this, we initially ranked the words in the dictionary in four tables: verbs, nouns, adjectives and particles. For example, the table Verbs contains the stem of the verb (verb in the past with the masculine singular person: $\text{he}$) and its primitive, but also its various semantic features, and its translation, Table 2.

**Table 2:** portion of the dictionaries.

<table>
<thead>
<tr>
<th>Verb</th>
<th>primitive</th>
<th>animated</th>
<th>translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>adjective</td>
<td>animated</td>
<td>translation</td>
</tr>
</tbody>
</table>

But during the implementation, we noted that there are special cases that must be treated separately.

Example: let the words: $\text{panel}$ = panneau
In fact, the analysis goes through three phases: a morpho-lexical analysis that aims to recognize each word in the sentence, a syntactic analysis to pull the various syntactic cases (subject, object...), and a semantic analysis. 

The system recognizes the action by comparing the words of the sentence with the verbs table, then consults the primary field (class of the verb) to extract the type of verb (ATrans, PTrans, ...), it makes an instantiation of this class and starts filling the fields. Then, the definition dictionary is consulted to extract the semantic features of each word of the sentence in order to analyze them:

- For example, the subject is recognized by the casual mark: Damma: 'ب', and according to the semantic features of the verb and nouns of the sentence.
- The instrument is recognized by the particles or words (بـ) and (بـ)،
- The source and destination with the particles: (بـ) and (بـ)
- Concerning the adjectives, and particles, they can be recognized easily (a table was devoted to the adjectives and another one, to the particles), etc.

At this stage, we get a representation of the sentence, independently of any language; this is what we call universal or internal representation. And it is in fact, the strength point of this approach, because it permits the generation of a translation to any target language; we just need to add a module to support it.

So, after the frame of the Arabic sentence (Arabic frame) is created, comes the role of the module: word for word translation, and we get the destination (target) frame. Then, the system dials the sentence from the target frame in a sequence that has been previously defined following the syntax rules of the target language (French), for instance:

Affirmative_Sentence = Agent Action [ ‘de’ + Source][ ‘à’ + Destination][ ‘avec’ + Instrument]...

The system and before providing the result, organizes the sentence following the rules of syntax and grammar of the target language (French):

- the time of Verbs: Present / past / future,
- the gender and number of names;
- All these treatments (sentence generation frame compliances of destination language) are provided by the module of management of the target language.

We note that the translation produced by the three phases: analysis, word for word translation and generation.

8 CONCLUSION & PERSPECTIVES

In this article is therefore, in the semantic processing of texts using purely linguistic and finds fulfillment with the DCF method as a basis. This method has
proved highly adaptable to the Arabic language and its peculiarities as to syntax and semantics [1], [4], [5], [6].

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

9 REFERENCES