MEANING-BASED TRANSLATION

Using a Semantic Analyser to resolve word sense disambiguation

Mohamed AZZEDINE
Managing Director CIMOS
73 Avenue Gambetta 75020 PARIS
Tel : 00 33 1 43 66 88 48 Email : azzedine@cimos.com

ABSTRACT
This paper describes a meaning-based approach for automatic translation. It is based on semantic analyser and lexico-semantic dictionary. The approach works at clause level and focus on universal grammar. Predicate and arguments are the nucleus and the periphery (number of arguments or valence). Topic/Comment/Tail are the unit of communication. They are called Clause at syntactic level and Proposition at semantic level. The ternary structure is universal and present at different level of the most of human languages.

KEYWORDS
Automatic Translation, Machine Translation, human Translation, Semantic Analyser,Word Sense Desambiguation, Lexical Analysis, Syntactic Analysis, Universal Grammar, Topic, Comment, Tail, Clause, Proposition, Predicate, Argument, Valence, Lexical-semantic dictionary.

"Translate meaning by meaning and not word by word"
Cicero

From its inception, CIMOS pursued two lines of business:
- professional translation services
- design/development of multilingual software (Machine Translation, linguistic tools...)

In response to the growing demand, we decided to **automate the translation process** for our own needs and then for those of our customers.

Machine Translation / Human Translation

Machine Translation (MT) software cannot translate the exact meaning of a sentence as well as a human translator. It approaches the exact meaning of the source text but will never reach all cultural connotations.
Human translators cannot hope to match the speed at which a computer can translate - at speeds of up to 20 words per second . Nor can human translators memorize and re-use all their previous translations (usually 2000 words per day is the expected productivity of a translator) the way a Translation Memory system does.

Translators and linguists would agree that there is no perfect translation because there are two types of translators. They can be either partisans of the source text or of the target text. Sometimes, the preference is given to one type of translator and sometimes it is given to the other. Another reason why there is no perfect translation is because the context and cultural environment create numerous ambiguities.

One has to accept imperfect but comprehensible machine translation output. Automatic Translation Software may offer a solution where speed is required or the volume of source documents is important. Machine Translation can provide comprehensible translations that are useful to the following scenarios:
- Translation of technical documents
- Translation for "gisting" in order to determine which documents to select for further, in-depth translation
- Quick turn-around translation at a reasonable cost

Translation background

The translation process of a document from one language to another language is a complex issue in itself. The automatization of this process is not easy to implement in MT software and more complex than human translation because the computer ignores nuances, connotations and implicit knowledge. How can we find for a source sentence an equivalent sentence in the target language?
Ancient Greek discovered that atom is the smallest part of physical matter. Around the same time syllabe and phonem were identified respectively as the smallest element of word and of distinctive speech sound.
Ibn Moqla who lived in 10th century (272-328 H) engineer and specialist in calligraphy added to these thruths: dot (or pixel) is the smallest part of an image
Sibawayh who lived in 7th century (172 H) linguist and specialist in arabic grammar added to these thruths that the smallest part of meaning is the three-pillars : Topic, Comment and Tail.
Topic and Comment define an independant meaning
which could be complemented by a Tail.
Tail is an adjunct like adverbial phrase (time or space) or prepositional phrase.
The unit of three-components Topic/Comment/Tail is called today Proposition at semantic level or Clause at syntactic level.

We have to split the meaning of a sentence in the smallest units that can exist in a given language. Obviously each of these smallest units has its own meaning.
The smallest unit that is common to all languages is the well known structure Verb, Subject and Object. These three components are central, and free word order. They are universal. This ternary structure is different from X-bar structure and it is used at different level (morphology, syntax, semantic…) in Arabic language. It is composed of a nucleus (*nawat*) with premodifier or prefix (*sabiq*) and postmodifier or suffix (*lahiq*).
Here is the word order for most of the human languages.

Free word order: 14 languages (Arabic, German, Greek...)

Fixed word order

SVO	SOV	VSO
55 languages	36 languages	16 languages
(English...)	(Japanese...)	(Spanish...)

Fixed word order

OSV	OVS	VOS
10 languages	3 languages	2 languages
(Amharic...)	(Balinese...)	(Italian...)

Now, the problem of Machine Translation can be reduced to the translation of the smallest units and the relationships between these units. How can we automatically parse sentence in its smallest units? Several approaches are present on the market and they all face the same problem: semantic disambiguation (or word sense disambiguation, WSD) because words are often polysemous.
The omission of vowels and diacritics (chedda, hamza), the free word order and the rich morphology of Arabic language increase the ambiguity of a source text.

Traditional Machine Translation approach

The most common approach involves the segmentation of the source text into words (multi-word), looking up the words (multi-word) in a dictionary (general and/or custom) , establishing the role of each word (multi-word) based on grammatical rules and building an equivalent of the source sentence in the target language..
The process includes the following steps:
- Morphological Analysis of the source text
- Syntactic Analysis of the source sentences
- Sentence Generation in the target language using morphological and grammatical rules

For example, let's take the Arabic sentence:
أعطي الولد تفاحة

أعطي the verb *"to give"* requires a subject and one or two objects
ولد the word *"child"* is not ambiguous
تفاحة the word *"apple"* is not ambiguous

For this simple sentence, the translation software should not have any difficulty using traditional electronic dictionaries to perform a straightforward look-up. The result is: *I give an apple to the child.*

Often, the bilingual electronic dictionaries are incomplete and give only some of the possible translations of a word. There is an artificial reduction of the number of possible target words due to the fact that the different meanings of a word exist within a specific context.

Meaning-based translation approach

The traditional analysis steps as previously outlined are not sufficient because they fail to identify the context and discover the deep structure of the sentence. In addition, they do not allow the resolution of ambiguity, especially in the case where ambiguity arises from polysemous words.

For example, with the new sentence
أعطى الرجل العلم يزيد

أعطي the verb *"to give"* requires a subject and one or two complements
رجل could be *"feet "* or *"man"*
علم could be *"flag "* or *"science"*

يزيدcould be verb" *add*" or surname "*Yazid*"

To resolve the ambiguity of this sentence, one must analyze the words in their context and identify the semantic links between the words. This is what is called semantic analysis. It includes the following steps:

1. Identify the predicate of the sentence which is the key element for building meaning.
2. Determine for each argument what its function and internal relationships are.
3. Use semantic information to resolve the ambiguity found in the sentence.

Let's go back to the sample sentence:
The verb *"to give"* requires between 1 and 3 arguments (valence)
A *"man"* can give a flag but a *"feet"* can't.
A *"man"* can't give *science* and neither can a *"feet"*.
Yazid can receive a concrete thing from a human but not from an inanimate object.

Obviously, to do the above, it is not enough to find the translation of the words. It is necessary to find the logical structure of the sentence within its own context and to have sufficient semantic information to help select the correct meaning.

The correct translation result is:
"The man gives a flag to Yazid"

The deep structure of the sentences takes into account contexts at different levels:
- the micro-context of the clause (semantic links)
- the local context of the sentence (interclausal relations)
- the global context of the paragraph (specific domain)
- the context of the document (cultural environment)

Extrapolating from the previous example, one can see that a large quantity of semantic information is needed to correctly translate a full text. For each word and for each concept found within a specific domain, the semantic information is made up of attributes such as human/animal; animated/unanimated; concrete/abstract countable/uncountable,... Additional information about the cultural context and world/ pragmatic knowledge are also useful. This is why it is necessary to build lexico-semantic dictionaries that are contextual.

Along with the semantic analysis steps, the meaning-based approach includes the tasks of clause identification and idiom identification.:
- Clause identification
For example the sentence:
أخذ الولد يلعب
has two predicates linked together
It translates as *"The child starts to play"* which is different from word by word translation *"The child took playing"*
- Identification of the idioms
For example the sentence
أخذ وأعطى معه
has a pre-defined meaning although it has a composite predicate

It translates as " *He deals with him* " which is different from word by word translation *"He took and gave with him"*
- Disambiguation uses semantic information attached to each word (for example, an unanimated concrete thing can't create an abstract thing ...). The disambiguation process operates beyond the level of the sentence to integrate the local context.

The enhanced approach for machine translation now includes the following steps:
- The Morphological Analyzer processes inflected words to find their roots and looks up the terms in the dictionary.
- The Grammatical Analyzer processes the arguments of the verb in the source sentences.
- The Clause Parser splits sentence in clauses
- The Semantic Analyser disambiguates and chooses the adequate meaning based on the context.
- The Sentence Generation arranges the words for the target sentence and outputs orthographically correct words using morphological rules.

Up until recently, semantic analysis was the forgotten step-child of natural language processing research and applications. With the focus on the semantic Web, one can find today some applications of semantics but still very few machine translation systems include semantic analysis.

Universal Case Grammar

The majority of the machine translation systems rely either on a statistically based approach (SMT), an example-based approach (EBMT), a rule-based approach (RBMT) or a combination of these three approaches.
CIMOS chose a meaning-based approach that makes use of the automated understanding of language. The syntactic analysis relies on the Universal Case Grammar theory.
Case Grammar theory first appeared in the Aristotle school and was further introduced in the Arabic grammar by the famous Arab linguist Sibawayh (7^{th} Century). However, this theory was never fully implemented and applied . A new approach for Arabic grammar will be published in 2005. It will embed the Arabic grammar with the Universal Case Grammar.
The following notions are universal:
- predicate, subject, object at syntactic level
- Topic, Comment; and Tail or *Musnad Ilayhi, Musnad and Fadlah* at semantic level
- relation between syntax (subject or *Fael*) and semantic (topic or *Musnad Ilayhi or Mubtada...)*

The Case Grammar Model has been recommended for use in natural language processing by C. Fillmore, (1968, 1971, 1977) and many other famous linguists. One can also find this grammar used in the Universal Network Language (UNL) project (see www.undl.org).
Case Grammar is not a traditional grammar. It deals with the semantic level of a grammar and within semantics it operates only with the inner structure of a single clause.

Information and word knowledge can be represented in terms of propositions that constitutes the basic units of communication. A proposition is a statement in which something is said about a subject. Each proposition is represented in syntax by a clause which holds a complete meaning. As a basic unit of communication, the proposition appears under different enunciative modalities in a discourse. The proposition may communicate assertion, interrogation or injunction.

The source text must first be segmented in clauses before the semantic analysis can start. This logical analysis is performed by the Clause Parser module that CIMOS has built for Arabic, English and French languages. The Clause Parser will identify: simple clauses, coordinate clauses, subordinate clauses, linked clauses....
The semantic analyzer developed by CIMOS also uses an ontology with 1,100 basic concepts. Each concept is described by semantic attributes.Each verb has a class such as state, activity, Kulub (heart verb), Chouruh (starting verb)....The semantic rules embedded in the analyzer represent a usage preference and not usage restriction.

Nakel Translator

The latest version of **NAKEL TRANSLATOR**, version 4.0, uses the **universal semantic analyzer** developed by CIMOS and for which CIMOS received the third innovation prize at LangTech 2003 in Paris. It is a bidirectional translation software for Arabic-English and Arabic-French language pairs. A multilingual version including the three languages is also available.

NAKEL software translates Arabic source texts for a given domain based on both dictionaries : general and specialized dictionary. In addition to the traditional steps of morphological and syntactic analysis, the software performs a semantic analysis before generating the target text.
It includes the ability to manage a user dictionary, by allowing the user to add new meanings.
It offers an interactive mode of translation in order to do proofreading with just a mouse click.

NAKEL software combines two independent processes in a sequential manner. First, the system performs a look-up in the Translation Memory (bilingual sentence database). If the search is successful, then the sentence is generated. Otherwise, if the search is not successful, it launches the Translation Engine (MT) which runs the different analysis steps and generation. The translation output may then be stored in the Translation Memory for re-use in future translations. The two processes are combined to improve the productivity of translators because it allows them to feed automatically the Translation Memory (TM) and avoid repeat translations.

Conclusion

Topic, Comment and Tail are the fundamental components of complex sentences in human language.

CIMOS's semantic analyzer is universal and independent of the syntax of a particular language because it operates at the level of the deep structure of the clause and uses semantic information and attributes described in a rich lexico-semantic dictionary combined with a set of concepts.
The use of lexico-semantic dictionaries and of a semantic analyzer that is built upon the Universal Case Grammar opens the way to a new enhanced machine translation approach .
The various ambiguities imposed by the complexity of the Arabic language (free word order, absence of vowels, ...) push to introduce statistical module to help resolving ambiguities at the final stage. This module is based on the frequency of meaning according to the word-use .
The meaning-based machine translation approach will significantly contribute to resolve word sense disambiguation. Machine translation will then translate a text meaning by meaning and get closer to simulating what a human translator does.

References:

1. Al Jahith (255 H) Al-Bayan wa at-Tabyin, Edition Haroun , T1 - 4 , Maktabat al-Hanji, Caro
2. Al-Jurjani, dalil al-Ijaz, Edition M.Rida, 1960, Caro
3. Allen , Jeff (1995). Natural Language Understanding
Second Edition. The Benjamin / Cummings Publishing Company
4. Ayoub G, La question de la phrase nominale en arabe littéraire: Predicats, Figures, Categories, T1 and 2 , Edition Septentrion, Presses Universitaires, Paris
5. Bohas G., Guillaume J.P. and D.E Kouloughli, (1990), The Arabic Linguistic Tradition, London-New-York
6. Chomsky, Noam (1971) Deep structure, Surface structure and Semantic Interpretation ; in Semantics, Edition Steinberg and Jakobovitz
7. Fillmore , Charles (1968) The case for case . E.Bach & R.Harms, Universals in Linguistic Theory
8. Guillaume J.P. (1986) Sibawayhi et l'énonciation, une proposition de lecture, Histoire, Epistémologie, Language, 7, II
9.Khayat M. (1996) Understanding Natural Arabic, in proceding KFUM workshop on information and computer science, Saudi Arabia
10. Sibawayhi, Abou Bichr, (192 H), Al-Kitab
Edition Haroun, Cairo 1977
11. Ibn Jinni (393 H) Al- Khasais, Edition Najjar, Beirut, T1 - 3
12. Strawson, P.F., (1974),Subject and Predicate in Logic and Grammar, London, Methuen & Co Ltd