نـــظـــام الـتـــرجـمـــة الآلـــيـــة
من العربية إلى الإنجليزية

MLTS

Multi-Lingual Translation System

Mohamed Azze-Edine*

From its inception, the CIMOS company pursued two lines of business: conventional translation services and design of a multilingual software. The translation process was carried out manually by skilled human translators trained on computer tools. The software developed by engineers with a strong linguistic background.

Over the year, our clients have submitted substantial documents to our translators such as technical manuals. In response to the growing demand, we decided to automate the translation process for our own needs and then for those of our customers. We created the department of "computer-assisted translation" with the hope of assisting better the translators. Its objective was to evolve the manual translation process toward a system of automated translation that could be marketed as a stand-alone product to our customers.

The MLTS software (Multi-Lingual Translation System) created in this department incorporates the lexical sources saved for more than 15 years of conventional translation services. MLTS is both a tool to help the translation process and an automated translation software with in an impressive level of performance on its own.

DESCRIPTION

MLTS was developed by a team of experienced linguists and software

* CIMOS PARIS FRANCE

engineers. It is an automated translation software conceived to help translators but not to replace them. It is capable of providing quickly the rough draft of an understandable and acceptable translation.

MLTS translates texts in a specific subject area (business, finance, computer science...) by using customized dictionaries.

The MLTS software operates on 5 different levels :

- translation memory
- morphological analysis
- syntactic analysis
- semantic analysis
- transfer.

MLTS combines the advantages of translation memory and of the "approach by transfer". it is based on semantic units and on an internal representation of languages concepts.

TRANSLATION MEMORY

The software starts by looking in the memory for an identical sentence to the one found in the text. It uses alignments techniques or similar approximation such as fuzzy matching. If the matching ratios is 100 %, there is an equality between two chains of characters. If the ratio is smaller, it is a sentence with subtle differences (declensions of nouns, pronouns, and adjectives, as well as verbal flexures).

The translation memory can suggest close correspondences (fuzzy matching). The software can carry out replacements of invariable parts (numbers, proper nouns) or variable parts with declensions.

The output of automated translation software can be viewed as o potential collection of initial translation memories. But this approach rapidly causes a saturation of the translation memory and increases research time.

Thus, we created new methods of arranging, matching, and compressing. For example, "give something to somebody". The new approach should recognize and intelligently generate adequate replacements according to the context.

- The woman bought a green dress.
- اشترت المراة لباسا أخضر
- The man bought a red suitcase.
- اشترى الرجل حقيبة سفر حمراء

The words "the woman" and "the man" are of the same category, the same goes for " a green dress" and "a red suitcase". Therefore, at the level of translation memory, these two sentences should be equivalent so only one from is saved.

MORPHOLOGICAL ANALYSIS

MLTS analyzes each word of the sentence in order to deduce grammatical characteristics (flexure, gender, number, person..)

It canonizes a word to its simplest form during a phase called "lemmatization". In other words:

- Nouns are brought back to their singular form
- Verbs to the infinitive form
- Adjectives to masculine singular
- Idioms to the infinitive form with pronouns in the 3rd person.

As a result, the software searches for the longest chain of words by comparing it to the elements saved in a collection of dictionaries :

- The general dictionary contains the usual words and locutions,
- The dictionary of idioms consists of expressions and locutions,
- The user dictionary consists of terms added or inserted by the user.

Each element of the sentence belongs to a linguistic category. This can be simple words, compound words or idiomatic expressions.

Sometimes, a given word has 2 or 3 potential translations falling into different categories. This type of ambiguity is frequent and additional procedures have to be set up to bypass this ambiguity.

For example :

- Samia put on her hat and left the room.

لبست سامية قبّعتها وغادرَت الغرفة

"left" can be : adjective, adverb, nom, verb

- Readers must confirm the information contained with other sources.

يجب على القرّاء أن يؤكّدوا المعلومات المُحتَوَاة في هذا النص مع المصادر الأخرى

"contained" can be : adjective, verb.

SYNTACTIC ANALYSIS

MLTS analyzes the structure of the sentence in order to recognize the grammatical categories of its constituents. Then the software verifies if the obtained structure conforms to the well-defined rules of syntax

MLTS analyzes each group of elements in the sentence to recognize the different constituents : verbal group, nominal group, and prepositional group. It identifies the different syntagms and builds the semantic links among the different constituents.

The syntactic analysis is generated (or generates) from the elements tagged with corresponding category and grouped in nominal group (NG), verbal group (VG), prepositional group -PG), etc... The determination of the subject group is a delicate problem and the software performs several iterations to solve it.

An additional problem to resolve is the notion of "heritage" between sentences. That is to say, the relationship among the word in different sentences. For example, the pronouns must agree in number and gender with their antecedents (nouns that they replace in the adjacent sentences).

- The Nile is a very long river, indeed it is one of the longest river in the world. It is longer than the Amazon.

النيل نهر طويل جدًا؛ حقًا إنه أطول الأنهار في العالم. إنه أطول من الأمازون.

"It" refers to the Nile. It is inherited from the preceding sentence.

- Arab universities graduate a large number of students each year. Some of these go to Europe or America to obtain advanced degrees.

تُخرّج الجامعات العَرَبية رقما كبيرا من الطلاب كلّ سنة. يذهب بعض هؤلاء إلى أوربا أو الولايات المتحدة الأمريكيّة لِيَحصُلُوا على شهادات عليا.

"these" refers to students. It is inherited from the preceding sentence.

SEMANTIC ANALYSIS

A sentence is an assembly of words expressing a complete thought.

It is composed of one or several clauses. A clause may be principal, independent, relative, coordinate, or subordinate, and is always governed by a verb.

The software determines the nature and the grammatical function of each element in the sentence, in addition to the function of the clauses that compose it.

The syntactic functions are reorganized according to their semantic role : predicate, arguments, and modifiers.

Often for a given word there are different synonyms which need to be sorted according to the context.

For example :

- We reached the village before darkness came on

وصلنا إلى القرية قبل أن يحلّ الظلام

Come on = حلّ أو تحسّن

- The rose is coming on well

تحسّنت الوردة جيّدا

Come on = حلّ أو تحسّن

TRANSFER

The transfer phase consists of two tasks : lexical transfer and structural transfer. Each sentence is represented by a tree which is diagrammed into the target language. The latter is particularly important in Arabic where the placement of the verb is before the subject at the beginning of the sentence. The progressive form of the English verb is either translated by an action noun (gerund) or by an incomplete verb.

- He is studying.

يدرس

- We are leaving tomorrow.

نحن مُسافرون غدا

GENERATION

The generation phase is a mirrored phase of the analysis phase. The operations of the morphological, syntactic and semantic analysis are performed in reverse order. The syntactic structure of each sentence of the source text is represented by a tree.

For each tree in the source language, a corresponding tree is built in the target language. Then, the syntactic and grammatical transfer is run in parallel. The difference between the two sentences is retained

At the same time, the software creates grammatical agreements (declensions of word) and verb

conjugations (flexures). This is the phase of morphological generation.

CONCLUSION

MLTS functions in batch mode or on-line.

- Batch mode : the resulting text is saved in a separate file from the source file
- On-line : the source text is translated sentence by sentence and the resulting text is displayed in a separate windows called Target Text, placed next to or below the Source Text window.

The unknown words are listed in a separate file which allows the user to enter them in the user dictionary.

The MLTS software works at about 20,000 words per hour and provides an understandable translation with great precision.

In the area of linguistic translation, the obtained results in a given stage are often far from being perfect (100% success does not exist) but they are understandable and acceptable. Even if they require improvements.

The chosen analyzer of English allows the creation of an automated translation software which offers a first draft translation warranting a post-edition phase.

Some questions raised at the syntactic and semantic level still remain to be resolved. In order to make up for these inconveniences, specifications for a new approach have recently been conceptualized and they will soon be implemented.

Some questions raised at the syntactic and semantic level still remain to be resolved. In order to make up for these inconveniences , specifications for a new approach have recently been conceptualized and they will soon be implemented.

We are making a special effort on the development of dictionaries :

- General dictionary (to have up to 100,000 entries)
- Specific dictionaries (Computer Science, Medicine, Finance, Business, Aviation,...)
- Dictionary of idiomatic expressions (to have up 10,000 entries).

MLTS OVERVIEW

	INPUT TEXT TRANSLATION MEMORY	1st step
ANALYSIS	INPUT TEXT (SOURCE LANGUAGE)	
	MORPHOLOGICAL ANALYSIS	
	SYNTACTIC ANALYSIS	
	SEMANTIC ANALYSIS	
	TRANSFER	2nd step
SYNTHESIS	SEMANTIC GENERATION	
	SYNTACTIC GENERATION	
	MORPHOLOGICAL GENERATION	
	OUTPUT TEXT (TRAGET LANGUAGE)	

SAMPLES OF TRANSLATION FROM ARABIC TO ENGLISH USING "MLTS" SOFTWARE

About the manuscripts

And the savants and the historians agree on that the history at the Moslems arose in the laptop of talk science, and the chest of Islam and till the third century of migration the science of talk and the science of historiography were merged, and the worker by them were named a speaker or a reporter, in other words a historian.
Wherefore many researchers consider that the science of historiography at the Arabs are really one of talk science branches.

The manuscript which was found in a small size and written by a beautiful naskhi calligraphy. The copyist emploies the red ink for the writing of headings and the black ink for the main part. And some possessions come under the heading from her " from books of the poor to the Most High God's mercy Ahmed ben Abraham ben Khalikan" and the manuscript includes sixty piece of papers the piece of paper number 20 was lost in her and some damage had caught up with it which did not impress his reading and his verification.

حول المخطوطات

ويجمع العلماء والمؤرخون على أن التاريخ عند المسلمين نشأ في أحضان علم الحديث، فمنذ صدر الإسلام وحتى القرن الثالث للهجرة كان علم الحديث وعلم التأريخ مندمجين، وأطلق على المشتغل بهما إسم محدث أو إخباري، بمعنى مؤرخ. ولهذا يعتبر كثير من الباحثين أن علم التأريخ عند العرب هو في الحقيقة فرع من فروع علم الحديث.

إن المخطوطة التي عثر عليها صغيرة الحجم ومكتوبة بخط نسخ جميل. استخدم الناسخ الحبر الأحمر لكتابة العناوين والحبر الأسود للمتن. وجاء تحت العنوان عدد من التملكات منها " من كتب الفقير إلى رحمة الله تعالى أحمد بن إبراهيم بن خلكان" وتحتوي المخطوطة على ستين ورقة ضاعت فيها الورقة رقم 20 وقد لحقها بعض التلف الذي لم يؤثر على قراءتها وتحقيقها.

SAMPLE OF TRANSLATION FROM

ENGLISH TO ARABIC USING "MLTS" SOFTWARE

▪ He traveled to Dubai, then returned to Bahrain.	■ سافر إلى دبي، ثم عاد إلى البحرين.
▪ The newspapers announced the result of the election.	■ أعلنت الصحف نتيجة الانتخاب.
▪ This program is not suitable for children.	■ ليس هذا البرنامج مناسبا للأطفال.
▪ Those women are skilled tailoresses.	■ أولئك النساء خياطات ماهرات.
▪ The government issues the official statistics at the beginning of the month.	■ تصدر الحكومة الإحصائيات الرسمية في أول الشهر.
▪ The national troupe will present a program of folk dance tomorrow.	■ الفرقة الوطنية ستقدم برنامج رقص شعبي غدا.
▪ Is your brother in the house?	■ هل أخوك في البيت؟
▪ How many cows are there in the field?	■ كم بقرة هناك في الحقل؟
▪ I happened to be in London at the time.	■ قد شاءت الظروف أن أكون في لندن في الوقت.
▪ It is going to be hotter.	■ إنه سيكون أسخن.
▪ It is now universally accepted that fresh water resources are limited.	■ إنه الآن مقبول بشكل شامل أن موارد الماء العذب محدودة.
▪ He is grateful to me for helping him.	■ يشكرني لمساعدته.
▪ The statement that he was going to resign surprised everyone.	■ إن البيان بأنه كان سيستقيل فاجأ الجميع.
▪ Readers are encouraged to confirm the information contained herein with other sources.	■ يشجع القراء لتأكيد المعلومات المحتواة في هذا النص بمصادر أخرى.
▪ Is the book useful ?	■ هل الكتاب مفيد؟
▪ Is there any mail for me?	■ هل هناك أي بريد لي؟
▪ Did you sell your car?	■ هل بعت سيارتك؟
▪ Whose book was she reading?	■ كتاب من كانت تقرأ؟
▪ How long did you stay in Paris?	■ كم مدة مكثت في باريس؟
▪ Have another cup of tea?	■ تريد كأسا آخر من الشاي.
▪ She put her dress on yesterday.	■ هي لبست لباسها البارحة.
▪ She get her dress today.	■ هي تتحصل على لباسها اليوم.

کتابخانه
بنیاد دایرة المعارف اسلامی

نحو توليد آلي للمصطلحات العربية

الدكتور / سعد بن خالد الجبري (*)

ملخص

تمتاز اللغة العربية كمثيلاتها السامية بقوة الاشتقاق الصرفي، وفي هذه الورقة نقدم تصورا نحسبه جديدا في نظرته إلى الدلالة الصرفية المبنية على الاشتقاق. ويعتمد هذا التصور على مستويين من الدلالة يمكن التمييز بينهما عند تحليل دلالة الاشتقاق، أولهما الدلالة المصاحبة للجذر والثاني الدلالة المصاحبة للصيغ الصرفية. إن تفاعل هذه المستويات الدلالية يولد مفاهيم متكاملة يمكن استيعابها على هيئة مشتقات عربية.

ولعل الهدف من هذه النظرة إلى الدلالة الصرفية يكمن في التعرف على سبل الربط بين المفاهيم الدلالية والكلمات العربية فيما تعورف عليه في علم توليد اللغة باسم الاختيار الصرف للمفردات Lexical choice. وتجدر الإشارة هنا إلى الطبيعة الخاصة للغة العربية كلغة اشتقاقية والتي تتطلب وسائل أخرى غير تلك وفرت للغات كالإنجليزية مثلا. ويعد الاختيار الصرفي للمفردات من الأمور اللازمة لأنظمة الترجمة الآلية ونظم توليد اللغة، ويتطلب تنفيذه توفر الوسائل اللازمة (لغوية كانت أم حاسوبية) للربط بين معنى معين وكلمة تستوعبه في لغة الهدف، وعليه فإن التوليد الآلي للمصطلحات العربية أقرب ما يكون في طبيعته إلى الاختيار الصرفي للمفردات نظرا لوحدة الهدف والمفهوم اللغوي والتقني.

وفي هذه الورقة تم التعرض وبشكل موجز إلى المواصفات الدلالية للمشتقات العربية ومن ضمنها المصطلحات إضافة إلى التقنيات المرتبطة لهذا المجال، وتم اقتراح تصور جديد يطال الاختيار الصرفي العربي المبني على دلالة الاشتقاق، والطرق الممكنة لتنفيذه آليا وذلك ببناء شبكة دلالية توارثية تؤمن التصنيف الآلي للمدخلات عند تنفيذها بواسطة أنظمة تمثيل المعارف وخاصة عائلة KL-ONE كما تم توضيح توجهات التصور الذي تقدمه هذه الورقة بأمثلة لمدخلات دلالية الهدف منها توليد بعض المصطلحات العربية، وجرى التنفيذ على مولد آلي يستخدم قاعدة معرفية مبنية على دلالة الاشتقاق في الصرف العربي.

(*) مدير مركز الحاسب الآلي الرياض