Overview of MT techniques

Malek Boualem
(FT)

This section presents an standard overview of general aspects related to machine translation with a description of different techniques: bilingual, transfer, interlingual and corpus-based technique including translation memory, statistical and example-based models.

1. Introduction

Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another. One of the very earliest pursuits in computer science, MT has proved to be an elusive goal, but today a number of systems are available which produce output which, if not perfect, is of sufficient quality to be useful for certain specific applications, usually in the domain of technical documentation. In addition, translation software packages which are designed primarily to assist the human translator in the production of translations are enjoying increasingly popularity within professional translation organizations. Comprehending the enormous complexity of translating human language and the inherent limitations of the current generation of translation programs is essential to understanding MT today [EAMT.99].

MT systems are designed according to one of the following parameters: coverage and reliability [CARL.99]. An MT system can either be designed to reproduce for a small language segment i.e. a sub-language or a controlled language with high fidelity and precision or it may be designed to perform informative, general purpose translations. In the former case, the system will have high reliability, whereas in the latter case, its coverage will be high. However, both properties are, to a certain extent, mutually exclusive.

- Coverage refers to the extent to which a great variety of source language texts can successfully be translated into the target language. A successful translation can be described as to be informative in the sense that allows a user to understand more or less the content of the source text.
- Reliability refers to the extent to which an MT system approaches an “ideal” translation (of a restricted domain) for a given purpose or for a given user. A reliable translation is user-oriented and correct with respect to text type, terminological preferences, personal style, etc.

2. Natural language ambiguities

One of the most important problems related to machine translation is the ambiguity of the natural language. We can distinguish several kinds of ambiguities:

2.1. Lexical ambiguities

This kind of ambiguity is related to homographs (words with identical transcription and different meanings). For example in the sentence: *Mark went to the bank*, the word “bank” can refer to a financial company or to the side of the river.

2.2. Semantic ambiguities

This kind of ambiguity occurs when a word (or group of words) can have different interpretations according to special contexts where it is (they are) used. For example in the sentence: *Mark may go by car*, it is not obvious to distinguish if the word "car" refers to a travelling way or to a specific car.

2.3. Referential ambiguities (anaphora)

This kind of ambiguity is related to referential structures (anaphora). For example in the sentence: *Mark took his car*, It is not obvious to know if Mark took his own car or another one.

2.4. Structural ambiguities

This kind of ambiguity occurs when a word (or a group of words) can have different syntactic values according to special contexts where it is (they are) used. For example in the sentence: *Mark went to work*, the word "work" can be considered either as a noun or as a verb.

2.5. Ambiguities related to idioms

The idioms are an important source of ambiguity. Unless they are identified as idioms, processing idiomatic expressions is very complex.

3. Generations of machine translation systems

Machine translation systems are classified into 3 generations:

3.1. Systems of the first generation

The main feature of these systems is the fact that different translation programs are designed for each couple of languages (bilingual translation). Moreover there is no separation between the programs and the linguistic data. They are based on linear (non-arborescent) data structures and do not use real computational linguistic methods such as regular languages or syntagmatic grammars. The *Systran* ancestor (developed at Georgetown University) is one of these systems.

3.2. Systems of the second generation

The main feature of these systems is that the translation process is developed into three different stages: analysis phase, transfer phase and generation phase. The analysis process transforms the source text into a source structural description which is transformed into a target structural description at the transfer phase and then to a target text at the generation phase. The systems of the second generation separate linguistic data (lexicons and grammars) from the processing programs. But these systems are not powerful at the semantic level. The first versions of the *Ariane* system (developed at GETA[1] Grenoble and used in the current UNL project) are classified in this generation of machine translation systems.

3.3. Systems of the third generation

The main feature of these systems is the ability to understand the meaning of a text before its translation. Due to the complexity of the natural language, these systems are generally dedicated to specialised and controlled languages. They are based on artificial intelligence techniques (expert systems and linguistic knowledge bases) to represent the semantic information of the texts. There exists no real functional system belonging to the third generation of MT systems.

[1] Groupe d'Etude pour la Traduction Automatique, IMAG, Université Joseph Fourier, Grenoble, France.

4. Translation categories

Translation is categorized into four types where a computer and a man can collaborate:

4.1. Machine aided human translation (MAHT)

The MAHT translation consists of using a word processing software completed by electronic dictionaries, which can be improved during the translation. Translations are human-made.

4.2. Human aided machine translation (HAMT)

This category of translation requires a human assistance before and after the automatic translation (pre-edition of the source text and post edition of the target text). The Canadian *Meteo* system is classified into this category of translation.

4.3. Interactive translation (IT)

In this category of translation, the system translates with an interactive human assistance. For each ambiguity problem during the translation process, the system asks for a human disambiguation. *Alps* is one of the interactive translation systems.

4.4. Machine translation (MT)

Theoretically the machine translation aims to completely avoid the human assistance to the system. Nowadays, no one of the existing machine translation systems can be qualified as being a MT system.

5. Machine translation techniques

In this section we introduce a descriptive presentation of the different machine translation techniques. These techniques are based on different models: bilingual, transfer, interlingual and corpus-based model which includes the memory-based, statistical-based and example-based models.

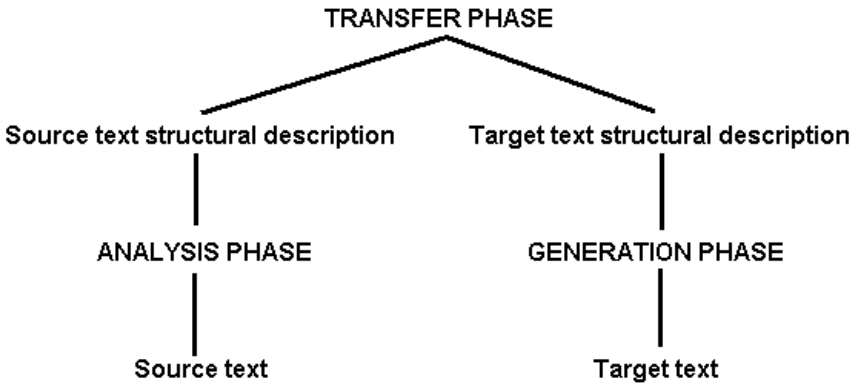
5.1. Bilingual-based machine translation

A bilingual machine translation system is dedicated only to a pair of languages and can not be adapted to other languages. Indeed the translation process is built according to specific characteristics of the two languages. A source text in one language is analysed to be specifically generated to another language. The transfer phase is minimised to bijective lexical and syntactical relations. It is understandable that the programs may be dependant on the language pairs making difficult their adaptation to new languages. The *Systran* system is a collection of bilingual sub-systems dedicated to different language pairs.

5.2. Transfer-based machine translation

The transfer translation model is built on three modules:

- Analysis module that transforms the source text into a source structural description.
- Transfer module that transforms the source structural description into a target one.
- Generation module that transforms the target structural description into a target text.

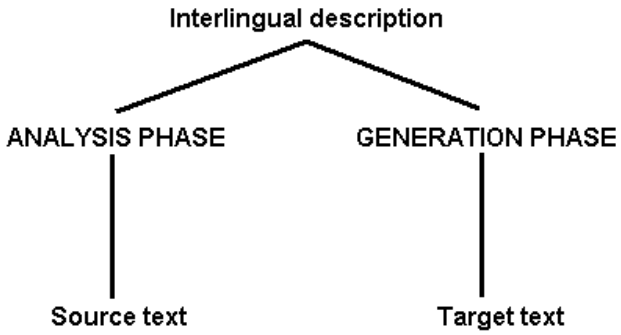


Eurotra European project system was built in the principle of the transfer model. It aimed to translate from/to the nine European languages. For each language, analysis, generation and transfer modules according to other languages have been developed..

5.3. Interlingual-based machine translation

The interlingual translation model is built on two main modules:

- Analysis module that transforms the source text into an interlingual description.
- Generation module that transforms the interlingual description into a target text.



The Ariane system developed by GETA research group is representative of this concept (pivot language) that is being used for the current UNL project.

5.4. Memory-based machine translation

Machine translation based on the “translation memory” is a corpus-based approach. It is dedicated to professionals or experts in the translation services. The system does not really analyses the source text to translate but just reuses possible translations previously stored by the professional translator. For the parts of text that have not been previously translated, a terminology (dictionary) support is used to help the expert to translate them. This “new” translation concept offers a computer-assisted translation that automates repetitive tasks, freeing the professional translator to attend to the finer points of translation that require the judgement of an expert.

The *IBM TranslationManager* [IBMTM.99] is one of the systems based on that concept. As an example, the following figure [IBMTM.98] shows the translation environment of IBM TranslationManager with the translation editor, the window for the translation memory proposals and the window for terminology support.

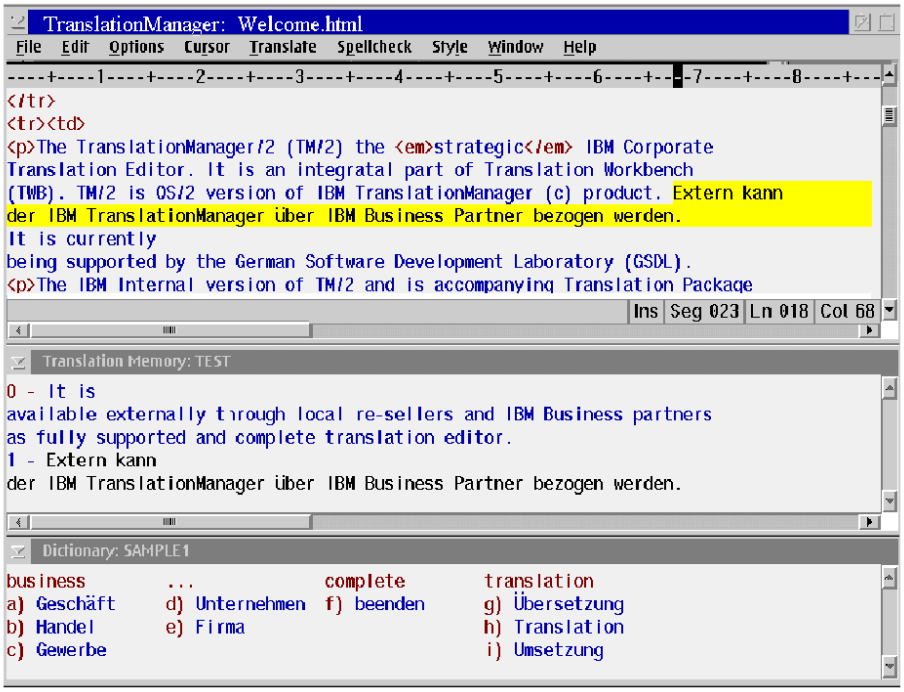


Figure extracted from the IBM web site
http://www.software.ibm.com/ad/translat/tm/tama/epaper.htm

Trados Translator's Workbench is a Translation Memory database which uses different translation editors/translation frontends for different file formats (http://www.trados.com/workbench).

5.5. Statistical-based machine translation

Statistical-based machine translation is a corpus-based approach. Statistical concepts are among the first techniques for machine translation. They was proposed by Warren Weaver in the early 1949's but that theories foundered on the rocky reality of the limited computer resources of the day. In the late 1980's IBM researchers felt that the increase in computer power made reasonable a new look at the applicability of statistical techniques to translation. Statistical machine translation was re-introduced by the *Candide* group at the *IBM Watson Research Center* [CAND.94]. The principle of this translation concept is that the computer inspects large collection of translated data and, from the collection, "learns" how to translate. However the statistical techniques are no longer encouraged in the machine translation domain.

5.6. Example-based machine translation

Example-based machine translation allows to rich systems. Translation examples are stored as feature annotated and sometimes structured representations. Translation templates are generated which contain (weighted) connections in those positions where the source language and the target language equivalences are strong. In the translation phase, a multi-layered mapping from the source language into the target language takes place. Sentences are more finely decomposed into phrases and linguistic constituents e.g. NPs, PPs, subject, object, etc. The example-based approach can make use of morphological knowledge and relies on word stems as a basis for translation. Translation templates are generalised from aligned sentences by substituting differences in sentence pairs with variables and leaving the identical sub-

strings unsubstituted. An iterative application of this method generates translation examples and translation templates which serve as the basis for an example-based MT system.

6. References

[EAMT.99] European Association for Machine Translation, http://www.eamt.org/

[IBMTM.99] IBM TranslationManager, http://www.software.ibm.com/ad/translat/tm/

[IBMTM.98] Overview of current IBM Translation Technology, TAMA'98, Vienna, 1998.

[CAND.94] A.L.Berger & al., The Candide system for machine translation, ARPA workshop on Human Language Technology, 1994.

[CARL.99] M.Carl, “Towards a Model of Competence for Corpus-Based Machine Translation”, Human Language Technology Center, Hong Kong University of Science and Technology, 1999.