# AN APPLICATION USING EXAMPLE-BASED MACHINE TRANSLATION

**Eliţa Natalia**
Ph.D Student
Technical University of Moldova
168, Stefan cel Mare bd.
Chisinau, Republic of Moldova, 2012,
*vnatalia@mail.md*

**Birladeanu Antonina**
Graduate student
Technical University of Moldova,
168, Stefan cel Mare bd.
Chisinau, Republic of Moldova, 2012
*toni_birlad@yahoo.com*

## Abstract

The aim of this paper was to study what an example-based translation is and explore the possibility to implement it in a system for translation of official documents. The language pair is Romanian-English and the language of the selected domain is relatively structured, which made work efficient.

## 1.Introduction

The basic idea in example based machine translation is quite simple: for the translation of a sentence, its previous translation examples are used.

A typical EBMT system is based on the following components:

- A database of aligned sentences in the source and target languages.The contents of the database, as well as its dimension are essential for the quality of selection. Examples have to be domain-relevant, long enough to capture specific particularities of a construction.
- A matching algorithm that identifies the examples that most closely resemble all or part of the input sentence
- A combination of algorithm which rebuilds the input sentence, through combination of retrieved fragments
- A transfer and composition algorithm that extracts corresponding target fragments and combines them into a sentence in the target language

## 2. Existing Example Based Machine Translation Systems

The evolution of EBMT systems could be observed by mentioning one of the first and one of the latest representatives.

One of the first EBMT systems is Eiichiro Sumita et al (1991, 1993), a system that translated only Japanese phrases of the form NOUN1 to NOUN2, where in the most contexts, the english translation was NOUN1 of NOUN2 and which used a commercial thesaurus of everyday japanese and calculated the semantic distance of the nouns, searching up the hierarchy for the most specific common abstraction.

HPA system, developed by Kenji Imamura (2001) includes the hierarchical alignment of the chunks which is functioning through determination of the equivalent chunks from the bilingual text depending on the correspondence of the words content and the same syntactical categories.

## 3. Developed system description

### 3.1. System architecture

For the translation of official documents the following system architecture was used as shown in Figure 1.

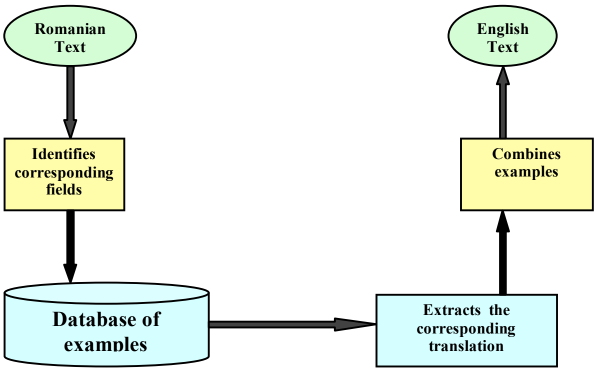

Romanian Text
English Text
Identifies corresponding fields
Combines examples
Database of examples
Extracts the corresponding translation

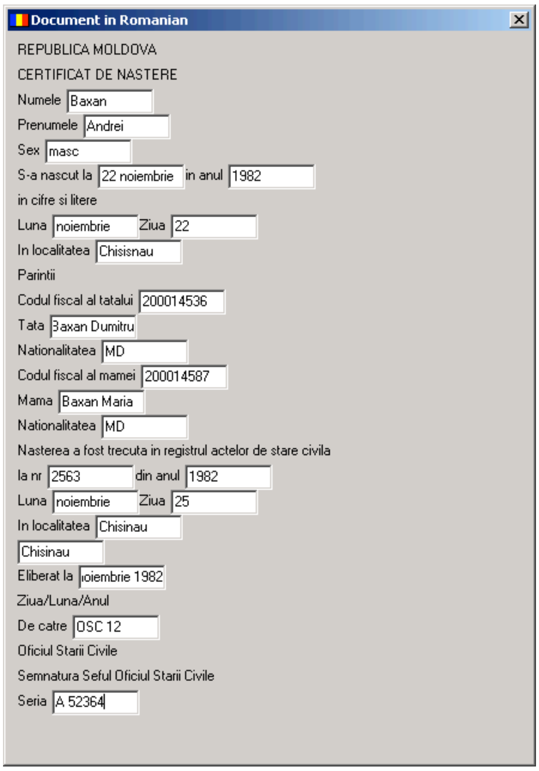
**Figure 1: System architecture**

The process begins by introducing a Romanian text. Then the system identifies the corresponding fields and consults the database of examples. Then the corresponding translation of the fields is returned and combined in an English text.

## 3.2 System Algorithm

### 3.2.1 Input data

As input data for this program separate files are used, which contain the document's texts. An example is presented in Figure 2. The user just has to fill in the fields with necessary information.



Document in Romanian

REPUBLICA MOLDOVA

CERTIFICAT DE NASTERE

Numele Baxan

Prenumele Andrei

Sex masc

S-a nascut la 22 noiembrie in anul 1982

in cifre si litere

Luna noiembrie Ziua 22

In localitatea Chisisnau

Parintii

Codul fiscal al tatalui 200014536

Tata Baxan Dumitru

Nationalitatea MD

Codul fiscal al mamei 200014587

Mama Baxan Maria

Nationalitatea MD

Nasterea a fost trecuta in registrul actelor de stare civila

la nr 2563 din anul 1982

Luna noiembrie Ziua 25

In localitatea Chisinau

Chisinau

Eliberat la noiembrie 1982

Ziua/Luna/Anul

De catre OSC 12

Oficiul Starii Civile

Semnatura Seful Oficiul Starii Civile

Seria A 52364

**Figure 2: User's interface of the translation tool. Input data for the Birth Certificate.**

### 3.2.2 The algorithm of the program

First of all we would like to mention that the English translation is created automatically, by searching field by filed in the database of examples the right translation of the field (phrase). By field or phrase we understand a sequence of words, that may not form a sentence, in our type of documents, such as birth certificates, marriage certificates etc, we cannot talk about full sentences, there are just certain sequences of words.

#### 3.2.2.1. Alignment

The alignment of corpora is manual. It is made at the field level, or in other words "filed by field". As a result we have an aligned database of examples, by linking the Romanian sequence of words with their equivalents in English: E.g *Numele si prenumele tatalui = Name and surname of the father*

#### 3.2.2.2. Translation

The translation of the fields (Figure 3) is made automatically by searching in the database of examples the appropriate translation of the field. In the moment you insert the data in the gap of the first field in Romanian, automatically the translation in the database of examples of this field is searched and displayed in the right part of the window. This translation is displayed row by row in the right part of the window. For example, we have the document "Birth Certificate", personal data is inserted in the Romanian document "s-a nascut la", then in the database of examples this field is searched. The document Birth Certificate is the fourth document in the database, and the field is 167 *"s-a nascut la" = born on.* But if we need the field Name, the filed would not be found in the rows reserved for the document "Birth Certificate", but in the row 53 reserved for the document "Work–record Card": *Nume=Name*, because it is repeated.

#### 3.2.2.3 Matching

In our project the matching algorithm includes the translation process. This statement is made because matching is made by mean of translation. For example we have the field"data nasterii" as an input data , then automatically this field is searched in the database of examples. When it is found it is imediatelly displayed on the window reserved for translation. The process of finding the right translation is considered to be the matching algorithm. So, matching algorithm includes matching the Romanian field with the English field of the future English document in the database of examples.

#### 3.2.2.4 Recombination

Generally speaking recombination at the sentence level is made for a more generalized example based machine translation system. First of all because this kind of system includes whole sentences and phrases to be translated. These sentences are divided into words, that are later searched and having been translated previously, must be recombined into a new sentence in target language. This is made because different languages have different morphology, different word order.

In our project we do not need this level recombination, because we do not deal with whole sentences. The only recombination that takes place is arrangement of all the necessary data found in the database of examples, in the right row according to the type of document to be translated. For this we design a file where the order in which the information in the translated document appears.

#### 3.2.2.1 Evaluation

Initially we would like to mention that this project was designed as a tool for translators. That is why firstly we have used for this project only 7 types of official documents. This tool also allows us to add some new documents to the existing one, edit the existing one, and to complete the database of examples. We cannot translate a type of document that is absent in our corpus.

Our database consists of about 1400 aligned examples, alignment being without any additional information, neither syntactic nor semantic. The system was tested for every type of document existent in our database, and because of the unambiguous database of examples and a controlled input, as could be seen from the above mentioned example, we estimate a rather good precision and recall result: P= 96 %, R= 6 %.

On the other hand, we cannot use the system without pre-defining a type of a document, as the input is controlled. So, the system is ineffective for new types of documents.

### 3.2.3 Output data

The output data are presented in a file „ *.txt ” , that contains the saved information.

An exemple could be presented in Figure 3.

REPUBLIC OF MOLDOVA
BIRTH CERTIFICATE
Name Baxan
Surname Andrei
Sex masc
Was born on 22 nd of november In year 1982
in numbers and letters
Month november Day 22
In the locality Chisinau
Parents
Father's Personal Code 200014536
Father Baxan Dumitru
Nationality MD
Mother's Personal Code 200014587
Mother Baxan Maria
Nationality MD
Birth was registered in the book of registration
No 2563 Year 1982
Month november Day 25
In the locality Chisinau
Chisinau
Issued on 27 november 1982
Day/Month/Year
by OSC 12
the registration office

Series A 52364

**Fig.3. The automaticaly generated English translation.**

### 3.3 Advantages and disadvantages

The following **advantages** were noticed:

- fast, effective, coherent translation
- unnecesity of having advanced knowledge of the user in English
- very effective tool for translators.

One of the biggest **disadvantages** observed is:

- a limited number of exemples and small corpora, so far.

## 4. Conclusions and further work

As a result of the study on what an example based machine translation is, how it is organised and our effort to create a system that uses EBMT principles, this tool to help translators emerged. We do know it is not perfect and needs a lot work to be improved, but it was the best we could do in a very short period of time, with limited human resorces.

The application has its limitations and this are the start point for further work. First of all we would extend the database of examples and complete it with syntactic and semantic features. As a further work too, can be mentioned adding a new direction ( English- Romanian) and possibly other languages. One more possibility is to apply it for another field.

## Bibliography

1. Brona Collins „Example-Based Machine Translation: An adaptation-Guided Retrieval Approach”,University of Dublin ,1998
2. Brown R.D.”Example Based Machine translation in the Pangloss system”,Pitsburg,1996
3. Chang-Baobao Zhang-Huarai Kang-Shiyong „Bilingual Corpus Construction and its Management for Chinese-English Machine Translation”, Peking University,1980,http://www.fi.muni.cz/usr/wong/teaching/mt/notes/mt.html
4. Chunyu Kit, Haihua Pan and Jonathan J.Webster „Example – Based Machine translation: A new Paradigm”, http://cslu.cse.ogi.edu/HLTsurvey/ch8node4.html
5. Cristina Vertan „Language Resources for the Semantic Web – perspectives for Machine Translation ”, 22527 Hamburg, Germany
6. Douglas Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys, Louisa Sadler, NCC Blackwell „Machine translation: An introductory Guide”, 1994, London, http://www.essex.ac.uk/linguistics/clmt//MTbook/HTML/
7. Eiichiro Sumita, Yasuhiro Akiba, Takao Doi, Andrew Finch „EBMT, SMT, Hybrid and More: ATR Spoken Language Translation System”, Kyoto, Japan, 2002
8. Federica Mandreoli, Paolo Tiberio „Searching Similar (Sub) Sentences for Example Based Machine Translation”, Universita di Modena e Reggio Emilia, Italy
iso-8859-1
9. John Hutchins „Machine translation: General Overview”, England,1992
10. John Hutchins „Research methods and systems designs in machine translation a ten-year review,1984-1994”,Cranfield University, England, November 2004
11. Kenji Imamura, Hideo Okuma, Taro Watanabe, Eiichiro Sumita „Example-based Machine Translation Based on Syntactic Transfer with Statistical Models”, Kenji Imamura, Hideo Okuma, Taro Watanabe, Eiichiro Sumita, Kyoto, 619-0288, Japan
12. Palmira Marrafa, Antonio Ribeiro „Quantitative Evaluation of Machine Translation Systems: Sentence Level”, Universidade de Lisboa, P-1050-050 Lisboa, Portugal