A first step in integrating an EBMT into the Semantic Web

Natalia ELITA
Informatics and Applied Modern Languages
Department
Technical University of Moldova,
168, Stefan cel Mare bd.
Chisinau, Republic of Moldova, 2012
vnatalia@mail.md

Antonina Birladeanu
Informatics and Applied Modern
Languages Department
Technical University of Moldova,
168, Stefan cel Mare bd.
Chisinau, Republic of Moldova,
2012
toni_birlad@yahoo.com

Abstract

In this paper we present the actions we made to prepare an EBMT system to be integrated into the Semantic Web. We also described briefly the developed EBMT tool for translators.

1 Introduction

The Semantic Web according to (Berners-Lee, 1998) is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.", The Semantic Web*,* Scientific American, May 2*001.* It provides a common framework that allows **data** to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming.

Further, we will present some general idea on what RDF is, then a general overview over EBMT, Some information on our system and the first step we made in order to integrate our product in the framework of the Semantic Web.

2 RDF notation

The Resource Definition Framework (RDF) is an entity relationship model used for representing information about resources in the World Wide Web. The main principle is that everything on the web can be unique identified with Uniforme Resource Identifier and then described in terms of triples representing the resources, their properties and values. For the purposes of the Semantic Web XML syntax was used, in this way the model benefits also from the Namespace propertyof XML and the RDF properties can be unique identified.

As the RDF properties can be organised in classes and subclasses, with attributes and values, it is often used to build ontologies. RDFS (McBride, Brickley, www) and other languages permit complete description of complicated ontological relatikons between RDF properties, in an RDF/XML format.

This is why we decided an ontology of terms used in an EBMT system could be the connecting element for the two areas: EBMT and the Semantic Web.

We will provide further a general overview over EBMT, then describe briefly the system we work on, at tne same time pointing to the ontology created and its place in the integration into Semantic Web..

3 General overview of the EBMT

The basic idea in example based machine translation (Brown, 2002) 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

It turned out that information about the syntactical structure of the fragments in both languages as well as pattern transfer rules, can improve significantly the performance of the example based MT system. Therefore it is quite usual that the example database contains, together with parallel aligned strings, also syntactic structures and their correspondences.

4 Developed system description

4.1 System architecture

For the translation of official documents the following system architecture was used (Fig 1).

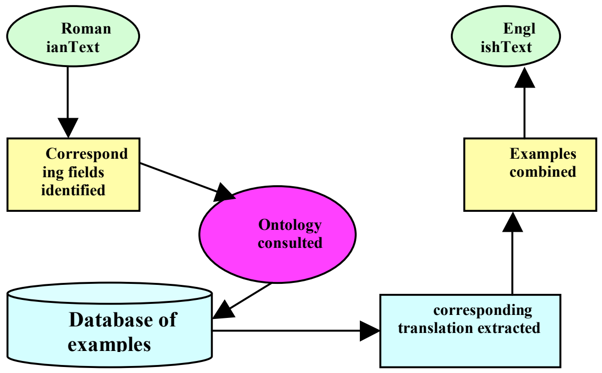

Fig 1. System architecture

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

We should mention here that Proteje 2000 was used to build the ontology of the terms, and we chose it, as we considered it to be the best tool, with the pre-defined RDFS option. Ontologies used in the translation systems have to be multilingual. That is different languages have to be mapped on it. This is why Proteje seems to be the best solution.

We will present further the basic characteristics of our system.

4.1.1 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.

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. Here are some examples:

Numele si prenumele tatalui = Name and surname of the father

Numele si prenumele mamei = Name and surname of the mother

Au incheiat casatoria in anul = Registered their marriage in the year

Casatoria a fost inregistrata in cartea de inregistrari cu nr. = The marriage was recorded in the book of registration Nr.

Locul inregistrarii actelor de stare civila = Place of the documents registration

Oficiul starii civile nr. = The office of registering official documents

Eliberat la = Issued on

Numele sefului Oficiului starii civile = Name of the official that made the registration.

Translation

The translation of the fields is made automatically by first consulting the ontology and then searching in the database of examples for 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

We consider that at this stage in our system Semantic Web resources should be used, as described in section 5 of the present paper.

Matching

At the moment, in our project the matching algorithm includes the translation process. This statement is made because matching is made by means 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 immediately 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.

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.

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. For adding a new document it is necessary to create a new "*.txt" file where the right order of information is indicated, and to complete the database of examples. We cannot translate a type of document that is absent in our corpus.

5 EBMT and the Semantic Web

As it was stated in (Vertan, 2003) the idea of the Semantic web implies a necessary standard annotation that is hopefully going to be used by the whole www community. RDF notation provides additional meaning, and machine translation, EBMT in particular, can make use of these additional annotations for several purposes. One of them was example-based rough translation. As RDF model aims at enriching documents with information about their content, it can help the example based rough translation. Also there, architecture for the extraction of translation correspondences was proposed. The organisation of the RDF annotation scheme is in two parts: syntactic and semantic. For our translation system we only need the semantic one as the syntactic information is not relevant in such documents as birth certificates, marriage certificates etc.

6 Conclusions and further work

In this paper we presented the first step made by us to prepare an EBMT system to be integrated into the Semantic Web. We also described briefly the developed EBMT tool.

The application has its limitations and these are the start point for further work. First of all we would extend the database of examples. 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.

Some further work should be done for the further integration of the sytem into the Semantic Web. So far, we only built the ontology of concepts, and we plan to use RDF to annotate all the documents in the corpus, and apply the techniques based on the suplimentary information provided by RDF for extraction of translation equivalents.

References

Douglas Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys, Louisa Sadler, NCC Blackwell,1994 Machine translation: An introductory GuideLondon

Tim Berners-Lee, 1998. Semantic Web Road map, http://www.w3.org/DesignIssues/Semantic.html

Brian McBride, Dan Brickley, RDF Vocabulary Description Language 1.0: RDF Schema http://www.w3.org/TR/rdf-schema/

Brown R.D.,1996, Example Based Machine translation in the Pangloss system ,Pitsburg,

Ralf Brown , 2002, Example Based Machine translation, Carnegie Mellon University

John Hutchins, 1992, Machine translation: General Overview, England,

John Hutchins, 2004, Research methods and systems designs in machine translation a ten-year review,1984-1994, Cranfield University, England,

Kenji Imamura, Hideo Okuma, Taro Watanabe, Eiichiro Sumita Example-based Machine Translation Based on Syntactic Transfer with Statistical Models, 619-0288, Japan

Palmira Marrafa, Antonio Ribeiro Quantitative Evaluation of Machine Translation Systems: Sentence Level, Universidade de Lisboa, P-1050-050 Lisboa, Portugal

Federica Mandreoli, Paolo Tiberio Searching Similar (Sub) Sentences for Example Based Machine Translation, Universita di Modena e Reggio Emilia, Italy

Eiichiro Sumita, Yasuhiro Akiba, Takao Doi, Andrew Finch, 2002, EBMT, SMT, Hybrid and More: ATR Spoken Language Translation System, Kyoto, Japan,

Cristina Vertan Language Resources for the Semantic Web – perspectives for Machine Translation, 22527 Hamburg, Germany