A First Step in Integrating EBMT into the Semantic Web

Natalia Elita , Antonina Birladeanu
Informatic and Applied Modern Language Dept.
Chisinau, Technical University of Moldova
vnatalia@mail.md
toni.birlad@yahoo.com

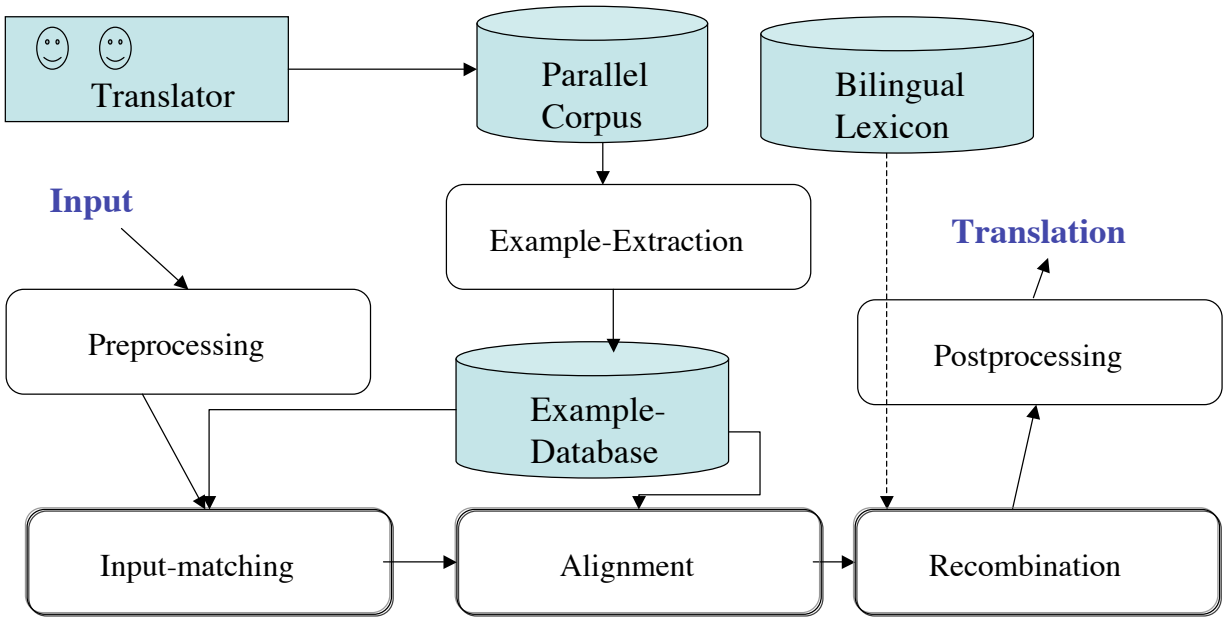
General Principles of EBMT

- A parallel corpus is used
- Part of the input text are compared with source chunks in the corpus
- The translation of the identified parts are put together and form the translation.

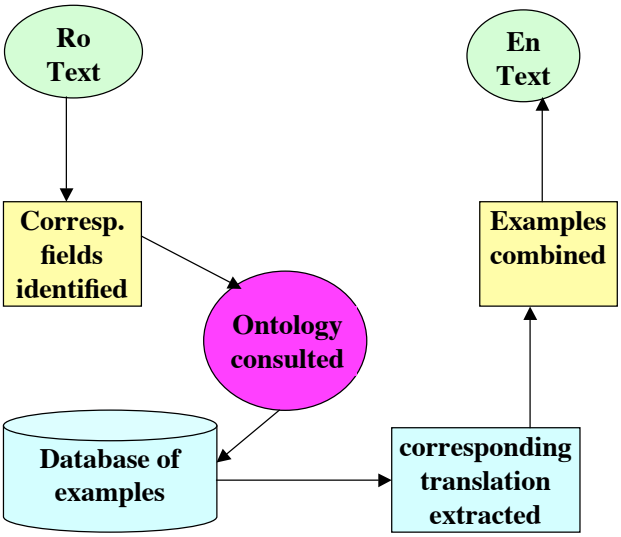
Functionality of an EBMT-System

- Relevant examples from a parallel corpus are extracted and saved in a database
- The input is compared with entries in the database(matching-phase).
  - Either the system looks for the identity of (parts of the) input with the database entries or
  - a distance between the input and the database entries is computed, and the database entry with the minimal distance to the input is chosen.
- Further on, in the alignement phase, the corresponding parts in the taget language are retrieved (this is trivial when the whole identical input is found in the DB)
- The corresponding chunks in the target language are recombined and build the output

Architecture of an EBMT-System



System architecture



Translation of legal documents

- Documents may vary across regions (in the same country) and time period but:
  - They have more or less the same structure
  - They use more or less the same vocabulary
- Syntax is quite simple, very simple sentences or, very common, parts of sentences.

System functionality -1-

- 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:
  - birth certificates,
  - marriage certificates etc,
- We cannot talk about full sentences, there are just certain sequences of words.

System functionality -2-

- 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.

Examples:

- *Numele si prenumele tatalui = Name and surname of the father*
- *Locul inregistrarii actelor de stare civila = Place of the documents registration*
- *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.*

Current development

- This project was designed as a tool for translators.
- 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.

Further work

- further integration of the sytem into the Semantic Web.
- At the moment only the ontology of concepts is built.
- Next steps:
  - 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.


Thank you