The EUCases Multilingual Access Module

Kiril Simov and Petya Osenova, IICT-BAS
Plan of the Talk

• Background
• Introduction of the EUCases Translation Module:
  – Ontology-Based One
  – Statistical Machine Translation One
  – Integration of both Modules
• Conclusions
• **WT3.5.** Development of multilingual access module.

• **Functionality.** The module translates a legislation query from a source language into a target language, and retrieves the detected texts that much the query.

• **Multilinguality.** The service is demonstrated for two languages – English and Bulgarian, in both directions (English-to-Bulgarian and Bulgarian-to-English).

• **Architecture.** The module consists of two submodules:
  – Ontology-based one, and
  – Statistical Machine Translation one.
Ontology-based Translation Module

• Relies on a common ontology with augmented lexicons
• The ontologies are: Syllabus ontology and Eurovoc multilingual taxonomy
• Exploits the ontology-to-text relation approach
  – The scenario consists of a (terminological) lexicon and a concept annotation grammar
  – The role of the lexicon is twofold: (1) interrelates the concepts in the ontology to the lexical knowledge used by the grammar for recognizing the role of the concepts in the text; (2) represents the main interface between the user and the ontology.
Ontology-to-Text Relation: Illustration
Ontology-to-Text Relation: Implementation

- It uses EuroVoc as a light-weight ontology
- EuroVoc provides aligned lexicons in all EU languages
- Since it covers the concepts related to the European Union activities, it is considered a domain ontology
- The concepts are arranged in domains and microthesauri
- The main relations encoded in EuroVoc that we exploit are: "skos:broader", "skos:related" and "xl:prefLabel", also its reverse relation "skos:narrower"
<table>
<thead>
<tr>
<th>Concept ID</th>
<th>Bulgarian term</th>
<th>English term</th>
<th>skos:narrower</th>
<th>skos:related</th>
</tr>
</thead>
<tbody>
<tr>
<td>1460</td>
<td>Финансов инструмент на Общността</td>
<td>EU financial instrument</td>
<td>1052, 2054, 2609</td>
<td>862, 1851, 2511, 4370, 5472</td>
</tr>
<tr>
<td>1052</td>
<td>Фондове на ЕС</td>
<td>EC fund</td>
<td>5138, 5643, 978</td>
<td>973, 4055, 8549, 862</td>
</tr>
<tr>
<td>5138</td>
<td>Структурни фондове</td>
<td>Structural funds</td>
<td>1056, 4056, 5668</td>
<td>5472, 5499, 5580, 5847</td>
</tr>
<tr>
<td>862</td>
<td>Помощ на Общността</td>
<td>EU aid</td>
<td>852</td>
<td>-</td>
</tr>
<tr>
<td>5499</td>
<td>Икономическо и социално взаимодействие</td>
<td>Economic and social cohesion</td>
<td>5864</td>
<td>5643</td>
</tr>
<tr>
<td>5643</td>
<td>Фонд за сближаване</td>
<td>Cohesion Fund</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Step 1: Text annotation
EU financial instrument ➔ <concept v="1460"/>

Step 2: Query expansion applying transitive closure of skos:narrower

<concept v="1460"/> ➔
<concept v="1460"/> <concept v="1052"/> <concept v="5138"/> <concept v="1056"/>
<concept v="4056"/> <concept v="5668"/> <concept v="980"/> <concept v="5643"/>
<concept v="978"/> <concept v="979"/> <concept v="2054"/> <concept v="2609"/>
Step 3: Query expansion applying skos:related

<concept v="1460"/> <concept v="1052"/> <concept v="5138"/> <concept v="1056"/>
<concept v="4056"/> <concept v="5668"/> <concept v="980"/> <concept v="5643"/>

→

<concept v="1460"/> <concept v="862"/> <concept v="1851"/> <concept v="2511"/>
<concept v="4370"/> <concept v="5472"/> <concept v="1052"/> <concept v="973"/>
Processing steps of the user query: EU financial instrument (3)

Step 4: *Deleting the repeated concepts*

Step 5: *Translation to the other language*

финансов инструмент на ЕС помощ на ЕС поддържащ механизъм земеделска валутна политика европейска парична система рамка за подкрепа на общността фондove (ЕС) европейски фонд за валутно сътрудничество европейски фонд за развитие европейски фонд за приспособяване към глобализацията структурни фондове икономическо и социално взаимодействие структурен разход подходящ район за развитие европейски фонд за регионално развитие регионална помощ регионално планиране регионална политика на ЕС
Ontology-to-Text Relation: Added Value

• The approach for *query expansion* is based on the intuition that when someone searches for a concept they are interested in all subconcepts of the given one as well as related concepts.

• Service options offered:
  – NQE: No query expansion
  – QNA: Query expansion using transitive closure of the relation `skos:narrower`
  – QRE: Query expansion using the relation `skos:relate`
  – QNR: Query expansion using both relations
Statistical Machine Translation Module

- As a system for statistical machine translation we selected Moses (a data-driven and state-of-the-art machine translation system).
- It provides the following translation models: phrase-based models; hierarchical or syntax-based models and factored translation models.
- The machine translation systems for the language pairs English-to-Bulgarian and Bulgarian-to-English were created using the Moses open source toolkit (see Koehn et. al. 2007).
- Parallel data from several sources was used to train factored translation models.
The parallel data

- SETimes (154K sentences)
  [http://opus.lingfil.uu.se/SETIMES.php](http://opus.lingfil.uu.se/SETIMES.php)
- Europarl (380K sentences)
  [http://www.statmt.org/europarl/](http://www.statmt.org/europarl/)
- JRC Acquis (364K sentences)
  [http://optima.jrc.it/Acquis/index_2.2.html](http://optima.jrc.it/Acquis/index_2.2.html)
- EAC-ECDC (7K sentences)
- APIS Legal Corpus (3844 sentences)
The monolingual data

- **Bulgarian**: National Reference Corpus (1.4M sentences) and the Bulgarian data from the parallel corpora
- **English**: Europarl (2M sentences) and the English data (without Europarl) from the parallel corpora
The data analysis

• We make use of the linguistic analyses produced by the Bulgarian pipeline btb-pipe (implemented partially within EUCases project) and the English system ixa-pipes (see Agirre et. al 2014)

• In the preprocessing step, we perform:
  – sentence splitting,
  – tokenization,
  – part-of-speech tagging,
  – lemmatization of the Bulgarian and English parallel corpora.
Translation Models (1)

- The SRI Language Modeling Toolkit (SRILM) Koehn et. al. (2007) was used to build 5-gram language models for the two translation systems.
- Two types of language models were used by both systems – word-form-based and part-of-speech-tags-based ones.
- Both systems were tuned with minimum error rate training (MERT) (see Och 2003) implementation provided as part of the Moses toolkit.
Translation Models (2)

• The **English-to-Bulgarian** translation system uses two types of linguistic factors - wordform and part-of-speech tag.

• The **Bulgarian-to-English** translation system uses three types of factors - wordform, lemma, and part of speech tag.

• Translation steps:
  – The input is processed with the corresponding pipeline, and factors are generated.
  – The factored data is lowercased (Moses script lowercase.perl).
  – The factored data is translated with the corresponding model.
  – The translation output is de-tokenized (Moses script detokenizer.perl).
  – The translation output is re-cased (a Moses re-caser model was trained for each translation direction).
Integration of Two Modules

• Why to combine?
  – The ontology-based translation module is not able to disambiguate ambiguous terms
  – It cannot translate out-of-vocabulary text
  – The SMT module is not able to translate words that are not mentioned in the training corpus

• Steps:
  – First, we translate the user query Q by each of the modules.
  – The results are Qot and Qsmt.
  – Each of them could contain substrings from the source language.
  – We delete them from the two translated queries. Then we concatenate the two strings.
  – The result is Qtrans.
  – This result is used for the full text search in a document database in the target language.
• **Linked Open Data** (Bizer et al. 2009)
  – Use URIs as (unique) names for things.
  – Use HTTP URIs so that people can look up those names.
  – When someone looks up a URI, provide useful information, using Web standards such as RDF, and SPARQL.
  – Include links to other URIs, so that they can discover more things.

  LOD is grouped in datasets, equipped with ontology, but this is not required by definition

• **Creation of a LOD dataset**
  – Selection of existing ontologies and dataset which defines the context of the dataset
  – Modeling of the current domain with respect to selected ontologies and datasets
  – Population of the dataset with instances
  – RDF repository: Loading (Inference) and Querying (SPARQL)
LOD Cloud (http://linkeddata.org/)

As of September 2011
EUCases Legal Linked Open Data

• EUCases Legal Linked Open Data
  – Facts extracted from case document: metadata and in text annotation
  – Ontologies: Dublin Core for representation of metadata, Syllabus ontology and Eurovoc multilingual taxonomy for representation of document annotation
  – Mapping to GeoNames
  – GraphDB of Ontotext for storing and querying

• Usages of EUCases Legal LOD
  – Searching for case documents using SPARQL
  – Searching for case documents via Multilingual access module and SPARQL templates
  – Definition of query expansion models
Conclusions

We have implemented two translation modules in an integrated way:
(1) statistical machine translation module and
(2) ontology-to-text translation module
The former exploits the progress in the area of machine translation.
The latter is based on a common ontology with aligned lexicons, and reflects the areas of query expansion and ontological inference.