Language Technologies in Bulgaria – Integrating Research with Innovations

Svetla Koeva

Institute for Bulgarian, Bulgarian Academy of Sciences
Sofia, Bulgaria
svetla@dcl.bas.bg

CESAR META-NET Roadshow
Belgrade, 29 October, 2012

Language technologies

- Language technologies (LT) are specialised information technologies for processing human languages in both modalities (spoken and written language) and in both directions (analysis and generation of language).

Language technologies survey

- The most comprehensive survey on the current state of LT in Bulgaria is provided recently by the White Paper Series.

Bulgarian Language White Paper

- In the category Speech Processing the quality of existing technologies for speech recognition and synthesis, the number and size of speech corpora and the amount and variety of speech-based applications are estimated.

Bulgarian Language White Paper

- Machine Translation: quality of existing MT technologies, number of language pairs, coverage of linguistic phenomena, amount and variety of available MT applications.

Вашата заявка:

Белград е голям промишлен център в Сърбия. Белград е университетски град и културен център. Белград е пристанище на две реки.

Превод:

Belgrade is a large industrial center in Serbia. Belgrade is university city and cultural business district. Belgrade is seaport in two rivers.

WebTrance

NLT research in Bulgaria

- Main areas of research and technology development: (i) the automatic translation of documents; (ii) the investigation of semantic models for linguistic description; (iii) large-scale information retrieval and content extraction.

MT research in Bulgaria

- A huge potential for improving the quality of MT:
  - Domain dependent lexical probabilities;
  - Effective word sense disambiguation;
  - Advanced morphology, syntax and semantic representations.

MT academic research

- SMT quality still obviously suffers from inaccurate lexical choice.
- This problem is addressed by integrating word sense disambiguation, where the WSD task is redefined to match not single words but word sequences for translation.

MT academic research

- In the traditional paradigm the MT relies on humans for a domain specification.
- The effective distinguishing between “better” translation hypotheses from their poorer alternatives is addressed by domain definition based on dynamic document categorisation.

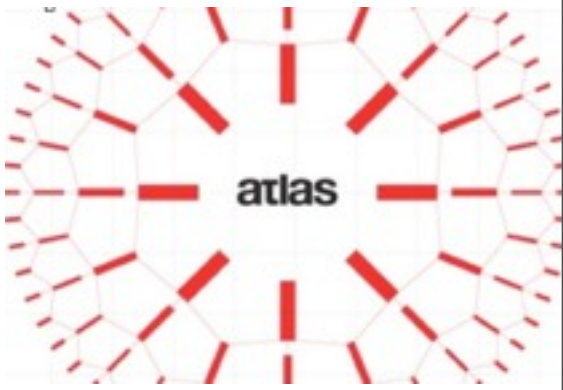
MT commercial research

- Chart parsing.
- Wordnet-like meanings from the initial charts are disambiguated with semantic rules.
- The target outcome is a complete syntactic parse with a wordnet sense assigned to each word in the sentence.

SkyCode

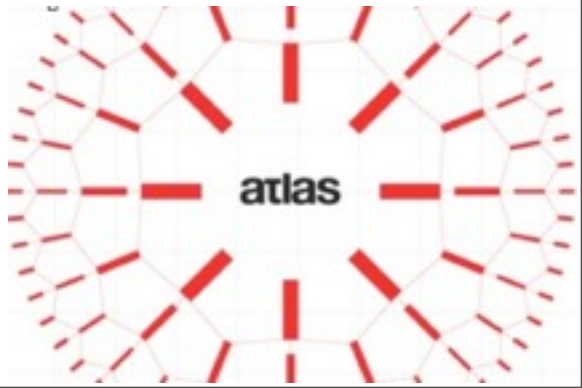
Research meets Innovations

- The academic and commercial research are met at the project ATLAS.
- Main purpose:
  - to facilitate multilingual web content development and management.
- Main innovation:
  - to integrate language technologies within a web content management system.



Research meets innovations

- Atlas provides automatic annotation of important words, phrases and named entities, suggestions for categorisation of documents, automatic summary generation, and machine translation of summaries of documents.



Research meets Innovations

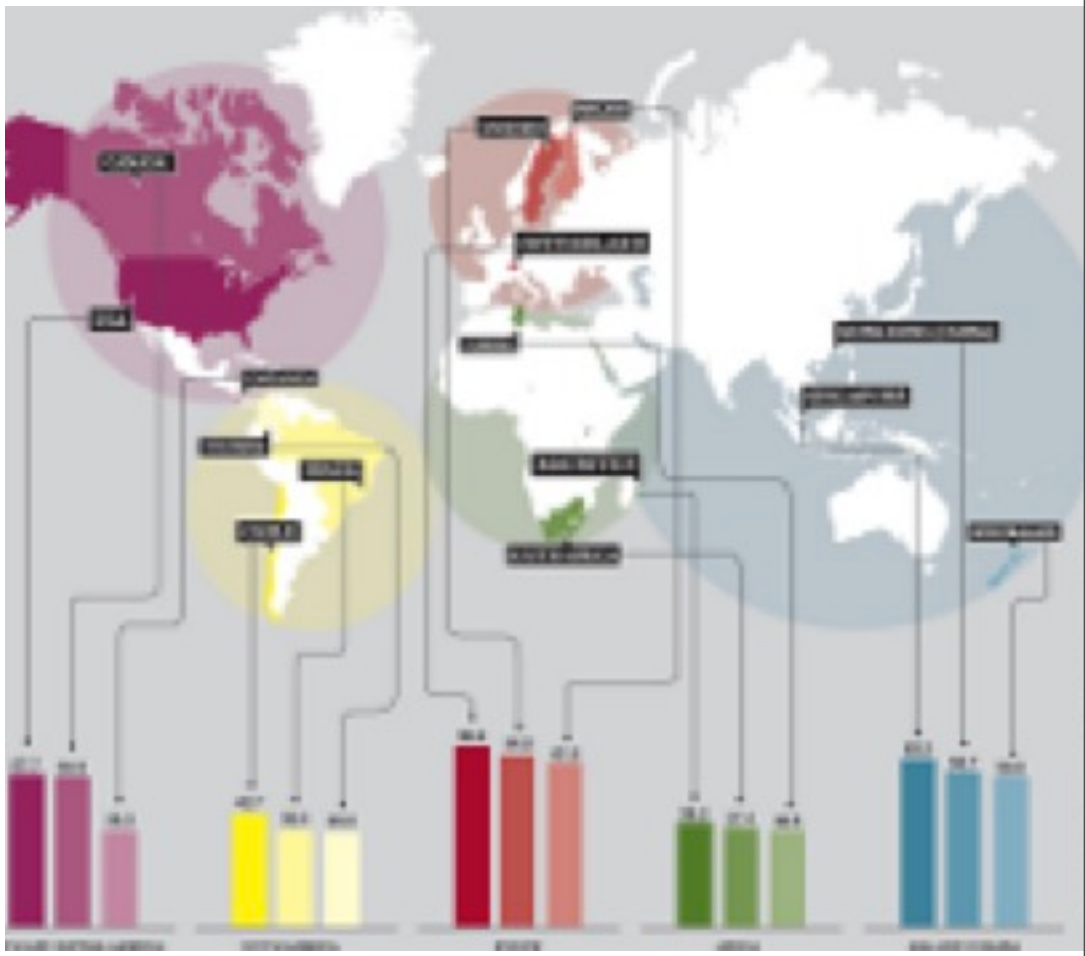
- To summarise, in the last years the research and commercial community on HLT in Bulgaria has expanded significantly, both for quantity and quality of its achievements.
- From the point of view of the methods, research is rapidly converging toward the use of empirical methods (i.e. based on data) while remaining semantically shallow.

Global Innovation Index 2012

- Global Innovation Index 2012: published by INSEAD, the leading international business school, and the World Intellectual Property Organization, a specialized agency of the United Nations.
- GII ranks 141 countries/economies on the basis of their innovation capabilities and results.

Global Innovation Index 2012 – top 10

- Switzerland
- Sweden
- Singapore
- Finland
- United Kingdom
- Netherlands
- Denmark
- Hong Kong (China)
- Ireland
- United States of America



Global Innovation Index 2012 – top 10

- The Report highlights a multi-speed Europe, with innovation leaders in Northern and Western Europe, Eastern European and Baltic countries catching-up, and a Southern Europe that performs less well.
- Bulgaria ranks 43rd. Croatia - 42nd, Montenegro - 45th, Serbia - 46th, Romania - 52nd and Macedonia - 62nd.

Global Innovation Efficiency Index 2012

- The Global Innovation Efficiency Index shows which countries are best in transforming given innovation inputs into innovation outputs.

Global Innovation Efficiency Index 2012

- Four of the top 10 countries in the Efficiency Index are 7 lower-middle income countries, Serbia being one of them.

- China
- India
- Republic of Moldova
- Malta
- Switzerland
- Paraguay
- **Serbia**
- Estonia
- Netherlands
- Sri Lanka

Conclusions

- There is a huge reasearch potential in the field of HLT.

Conclusions

- Research and innovation open up numerous new business opportunities for European language-technology and -service providers.