ARTIFICIAL INTELLIGENCE HANDLING TEXT DATA

Dunja Mladenić, Marko Grobelnik
J. Stefan Institute, Slovenia

ailab.ijs.si

Word: sound – written – analyzed

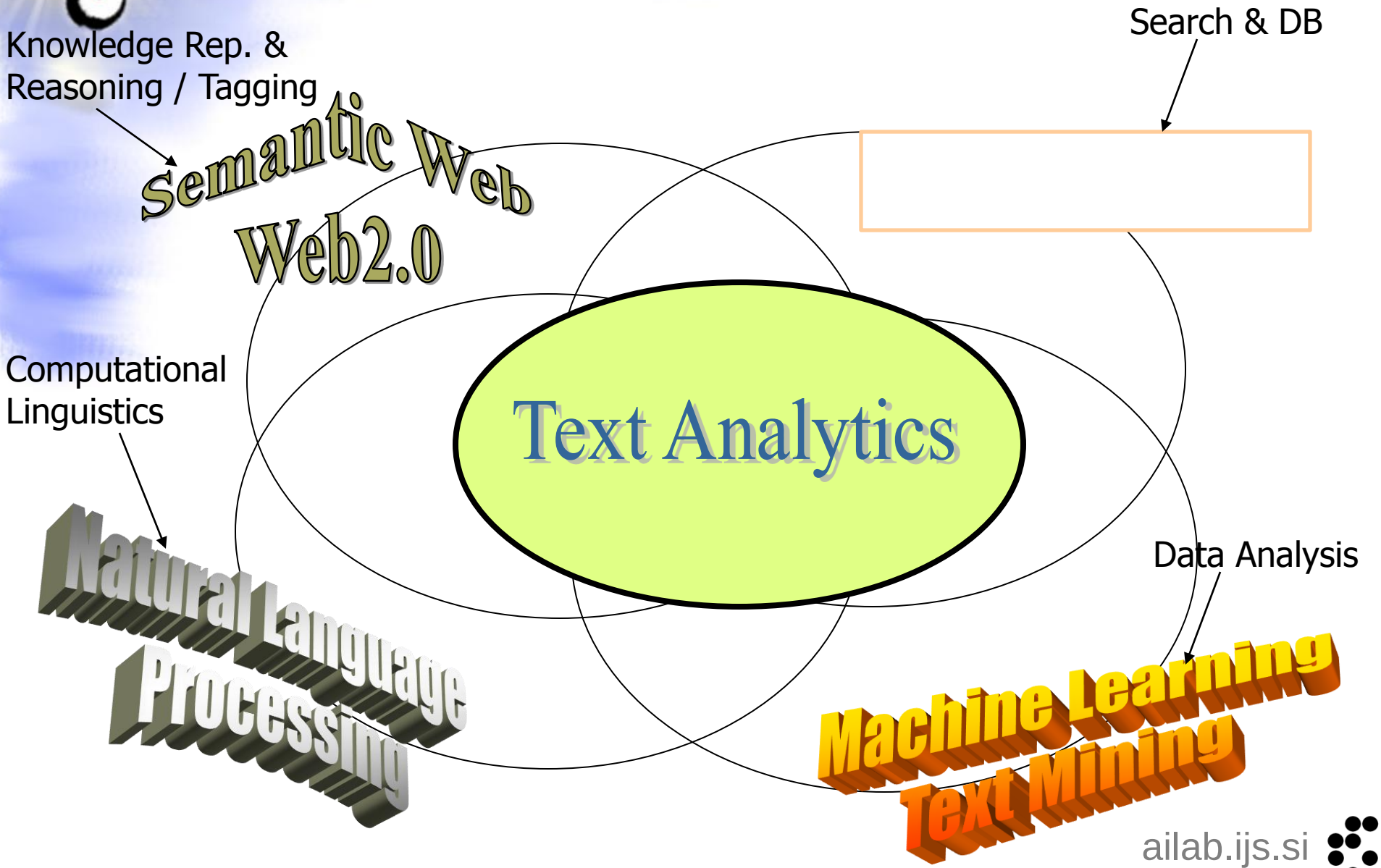
"Every element of the universe is in a constant state of vibration manifested to us as light, sounds and energy. The human senses perceive only a fraction of the infinite range of vibration, so it is difficult to comprehend that the **Word** mentioned in the Bible **is actually the totality of vibration which underlines and sustains the creation**."

[Y. Bhajan]

Outline

- Text for human readers but also to be handled by computers:
  - editing, storing and indexing, searching and retrieving, ranking, classifying, extracting information and knowledge, question answering,…
- AI research in Slovenia involving handling text
  - Research problems
  - Example tasks with demos

Handling text data



Text Learning in 1990s

- Machine learning methods and tasks
  - Inspired by information retrieval, classifying document regarding relevance for a query
  - Personalized information delivery
  - Document categorization into class hierarchy

Personal WebWatcher [Mladenić, 1996]

Yahoo Planet [Mladenić, 1998]

Carnegie Mellon University, Pittsburgh, 1996

Machine Learning on Text Data 1990s

- Text representation
  - Words and word sequences as features in ML setting
  - Handling sparse feature vectors
  - Efficient feature selection

Word sequences
[Mladenić and Grobelnik, 1998]

Feature selection on hierarchy
[Mladenić and Grobelnik, 2003]

Ljubljana, 1995

Ljubljana, 1998

Text Analytics in 2000s

- Information extraction from the Web
- Combining text, graphs, databases, images,…
- Capturing semantics with Cyc
- Knowledge management

Mining symbolic knowledge
[Ghani, Jones, Mladenić, Nigam, Slattery, 2000]

Text mining and link analysis
[Grobelnik, Mladenić, 2003]

Text mining and link analysis
[Baxter, Klimt, Grobelnik, Schneider, Witbrock, Mladenić, 2009]

Automated knowledge discovery
[Grobelnik, Mladenić, 2005]




Example Tasks

- Visualization of text data
- Semi-automatic ontology construction
- Text annotation
- Extracting triplets from text
- Document summarization
- Question answering
- Ontology construction and extension
- Social network analysis and text analytics

Visualization of news [Grobelnik, Mladenić, 2004]

OntoGen [Fortuna, Grobelnik, Mladenić, 2005]

Enrycher [Štajner, Rusu, Dali,Fortuna, Mladenić, Grobelnik, 2010]

Semanitc Graph [Rusu, Fortuna, Grobelnik, Mladenić, 2009]

Document Semanitc Graph [Leskovec, g, Grobelnik, 2004]

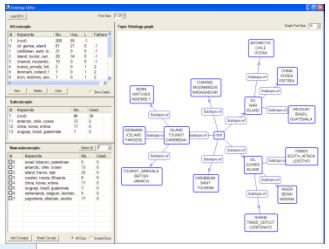
AnswerArt [Dali, Rusu, Mladenić, 2009]

OntoPlus [Novalija, Mladenić, 2010]

Semanitc Spaces [Fortuna, Mladenić, Grobelnik 2009]

Technologies 2012

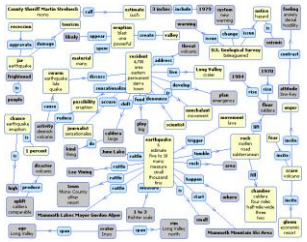
Deep Semantics & Reasoning (Cyc)



Light-Weight Semantic Technologies (OntoGen)

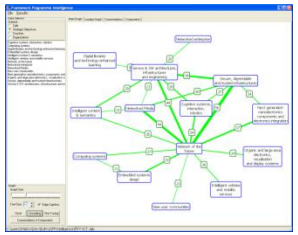
Social Computing/Web2.0 (LiveNetLife)

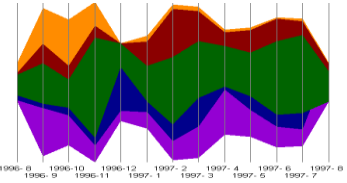
Computational Linguistics
(Enrycher, AnswerArt)



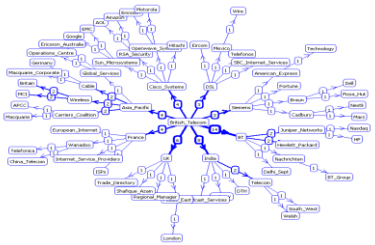
Complex Data Visualization
(DocAtlas, NewsExplorer, SearchPoint)

Graph/Social Network Analysis
(GraphGarden/SNAP, IST-World, FPIntelligence)





Data/Web/Text/Stream-Mining
(TextGarden Suite of tools)



Statistical Machine Learning

videolectures.net
exchange ideas & share knowledge

ailab.ijs.si

Discussion

- AI successfully applied on text data
- Number of technologies and prototypes developed
- More sophisticated methods result in addressing more complex problems
- On going research:
  - Big data including stream of text
  - Cross-lingual, cross-modal, cross-domain