Adaptive Semantic Exploration of Historical Textbook Corpora
Carsten Schnober (UKP-DIPF)

Context
• Supervised by Prof. Dr. Gurevych
• Takes place within the project Welt der Kinder, funded by Leibniz Association:
  • First work that is based on a newly digitized corpus of historical schoolbooks from 1850-1918
  • comprising 7,000 books with 600,000 pages; continuously being extended
• Project Goals:
  • Explore representations and interpretations of the world
  • Explore the particular media type
  • Combine established hermeneutic methods with text mining and scientific annotation technologies
  ▶ Computational linguistics research for opinion mining on special texts without existing (annotated) training data
• Cooperation between nine German and Swiss institutes and universities
• Project duration: 01/05/2014 – 31/05/2017

Research and Goals
• Support historical and linguistics research ➔ Digital Humanities
• Challenges:
  • Corpus language is particular and diverse
  • Erroneous text due to OCR
  • Humanities researchers cannot define specific information needs, since they are not familiar with text analysis
• Unsupervised and weakly supervised learning:
  • Infer information from the data with little or without prior knowledge
  • Adapt existing methods from topic modelling, opinion mining, sentiment analysis
• Develop new approaches to gain information valuable for humanities researchers

Methodology and First Steps
• Iterative collaboration with project partners: cycle of development, prototypical applications, feedback and further development
• Topic Modelling (LDA, Blei et al., 2003) (Figure 1)
• Joint Topic-Semantic Modelling (JST, Lin & He, 2009)
• Artificial neural networks and deep learning (Bengio et al., 2006)
• Visualize data and results of text analysis

References

Figure 1: Topics and document assignments with LDA (Blei et al., 2003)