Preparing Multi-Modal Data for Natural Language Processing

Erik Novak, Jasna Urbančič, Miha Jenko
Jožef Stefan Institute
Ljubljana, Slovenia
Introduction

• Students and teachers are searching learning materials for their education

• Millions of education material found
  • Multiple modalities (text, video, audio, etc.)
  • Different languages
  • Different learning preferences

• Pre-processing pipeline that handles multi-modal and cross-lingual data

• Solution can be applied on other domains
Outline

• Pre-processing Pipeline
  • Crawling
  • Formatting
  • Text Extraction
  • Wikification

• Data Statistics

• Application: Recommender Engine
Pre-processing Pipeline
Pre-processing Pipeline

Crawling

• Targeted four OER repositories
  • MIT OpenCourseWare
  • Università di Bologna
  • Université de Nantes
  • Videolectures.NET

• Used dedicated APIs and custom crawlers

• Acquired material metadata
  • title, description, url, type, language, provider
Pre-processing Pipeline

Formatting

• Designate which material attributes are required
• Setting up a schema for checking missing material attributes
Pre-processing Pipeline

Text Extraction

• Extracting content from the material in text form
• Handle each file type separately
  • Text – *textract*
  • Video and audio – *transLectures*
Pre-processing Pipeline

Wikification

• Linking material textual components to the corresponding Wikipedia page

• **Wikifier Service**
  • Finds Wikipedia concepts that are related to the textual input
  • Supports cross- and multi-linguality
  • Input text is limited to 20k characters
Data Statistics

- Acquired and pre-processed approx. 90k items
- Repositories covering 103 languages
  - Graph showing languages with at least 100 materials
Data Statistics (cont.)

• Each file type can be represented in various formats

• Most dominant type – text (pdf, pptx and docx)
  • Followed by video (mp4)
Application: Recommender Engine

Content-based recommender engine

- Using $k$-nearest neighbour algorithm
- Comparing materials’ Wikipedia concepts – giving cross-lingual recommendations
- Wikipedia concepts extracted from material content – providing multi-modal results
Conclusion

• Methodology for processing multi-modal and cross-lingual items

Future Work

• Improve text extraction methods
• Handle missing material attributes
• Add new feature extraction methods to determine quality and topic of material

Acknowledgements: This project was supported by the Slovenian Research Agency and the X5GON European Union’s Horizon project under grant agreement No 761758.

Icon made by Freepik from www.flaticon.com