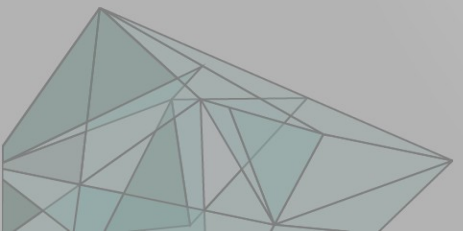
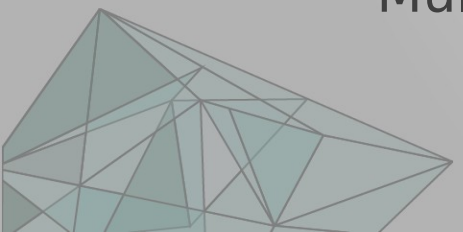
ExB Text Summarizer

Stefan Thomas, Christian Beutenmüller, Xose de la Puente,
Robert Remus & Stefan Bordag
ExB Research & Development GmbH



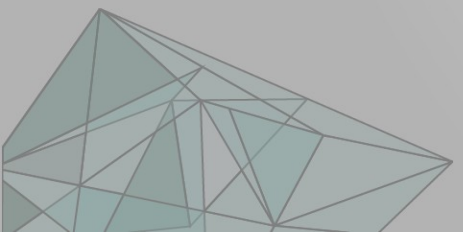
Outline

- Introduction
- ExB Summarizer
  - Preprocessing
  - (Fair)TextRank
- Official results
  - Single document summarization
  - Multi document summarization



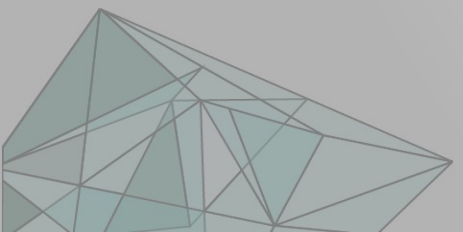
Approaches to summarization

- Sentence extractive methods
  - State of the art
  - Limited results
- Abstractive summarization
  - The “human way“ of summarization
  - Additional difficulty: Producing text



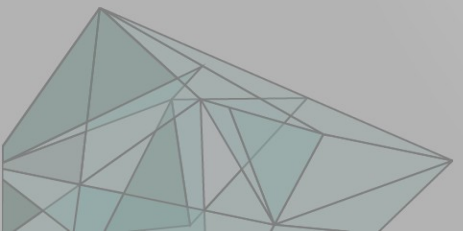
Overview of ExB summarizer

- Goals:
  - Scalability
  - Language independence
- Key ideas:
  - Main parts unsupervised
  - TextRank on a similarity graph of sentences



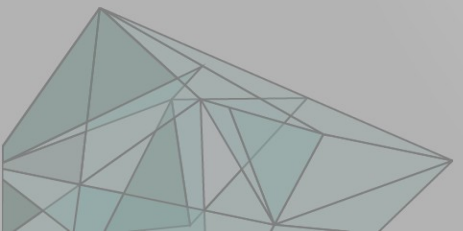
Preprocessing steps

- Rule-based Tokenization
- Stop-word removal
- Stemming
- Sentence boundary detection
- Temporal expression detection for multi document summarization



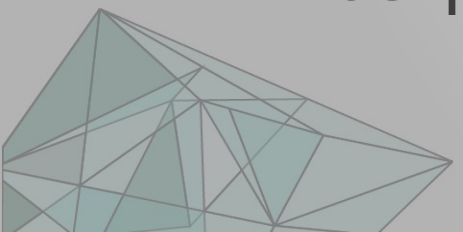
TextRank

- Invented by Rada Mihalcea
- Origin: PageRank algorithm [Page & Brin]
- Graph-based ranking algorithm
  - Text is represented as nodes and edges
  - Nodes are ranked according to their importance



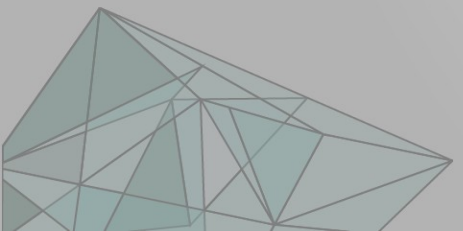
FairTextRank

- Sentence similarity graph
  - Sentences = nodes
  - Similarity between sentences = weighted edges (between 0 and 1)
  - Bag-of-words model with Jaccard index
- Iterative application of TextRank
  - Helps covering different topics in the produced summary
- Postprocessing



Results (MSS)

System	#langs	Rank R-1	Rank R-2	Rank R-3	Rank R-4	Rank R-4SU
BGU-SCE-M	3	2.0	3.3	3.7	4.3	3.0
BGU-SCE-P	3	5.0	4.7	5.0	4.3	4.3
CCS	38	2.1	2.1	2.2	2.3	2.5
ExB	38	3.2	3.3	3.7	3.8	2.8
LCS-IESI	38	4.1	4.1	4.0	4.0	4.1
NTNU	2	5.5	6.0	6.0	7.0	5.0
UA-DLSI	3	6.0	5.0	4.7	5.0	6.0
Lead	38	5.1	5.0	4.6	4.3	5.0
Oracles	38	1.1	1.2	1.2	1.2	1.2

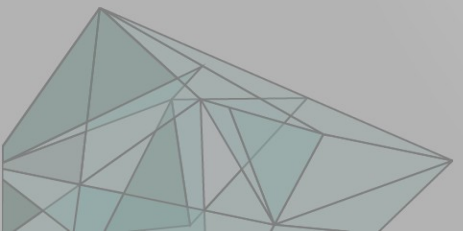


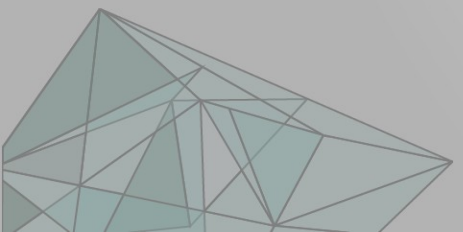
Results (MMS)

Language	AutoSummENG	MeMoG	NPowER	Rank/Total
Arabic	0.135	0.164	1.717	7/9
Chinese	0.118	0.141	1.654	1/5
Czech	0.188	0.200	1.874	4/7
English	0.167	0.191	1.817	6/10
French	0.200	0.195	1.892	5/8
Greek	0.147	0.170	1.750	5/8
Hebrew	0.115	0.147	1.655	8/9
Hindi	0.123	0.139	1.662	3/7
Romanian	0.168	0.183	1.809	4/6
Spanish	0.193	0.202	1.886	3/6

Summary

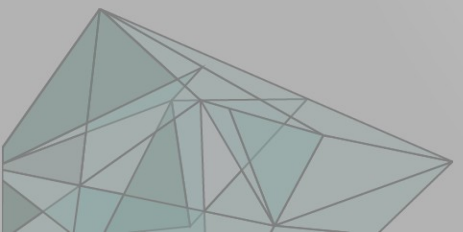
- TextRank based approach
- Multitude of preprocessing steps
- Participated in all possible languages
- Competitive results



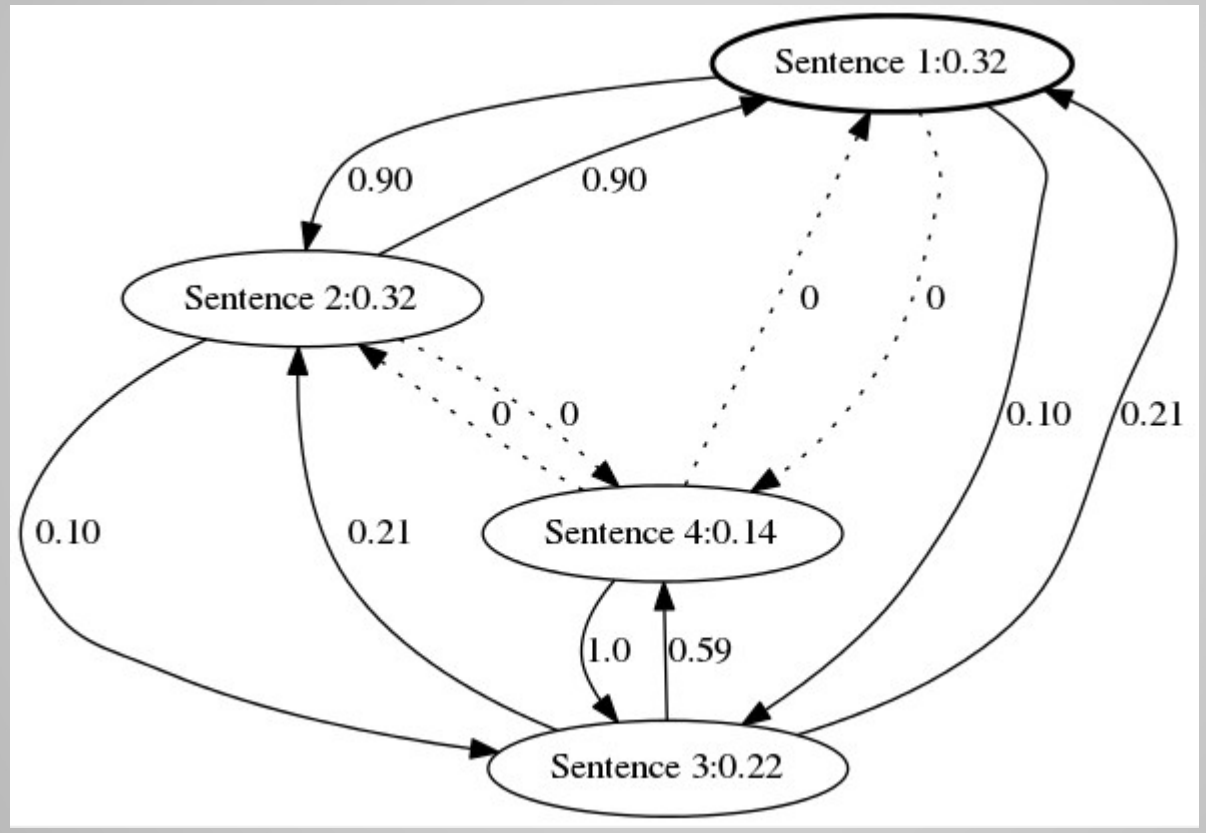


Negative findings

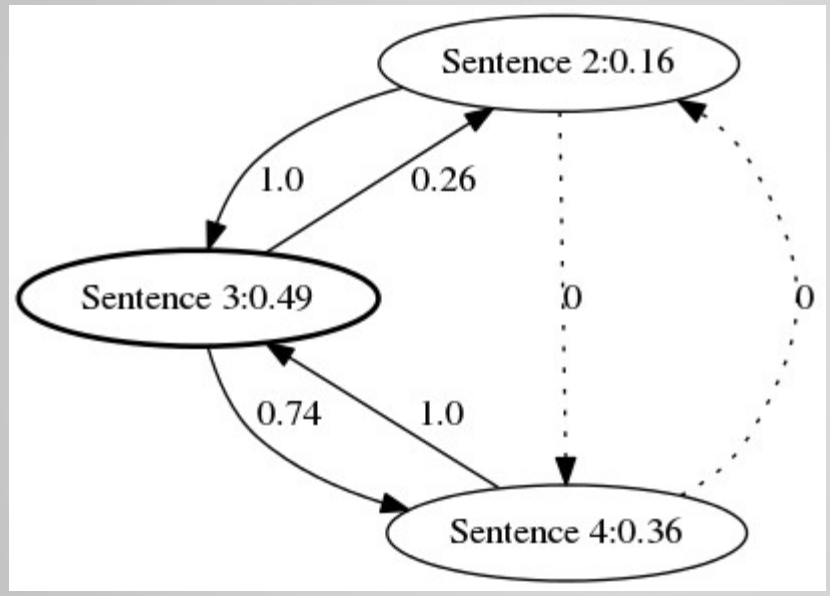
- No increase in performance via:
  - Semantic Text Similarity instead of bag-of-words/Jaccard index
  - Word2vec word embeddings
  - Named entities
- ROUGE measure is inappropriate



Example of FairTextRank



Example of FairTextRank



Example of FairTextRank