Using Content and Structure at WiQA 2006

Sisay Fissaha Adafre, Valentin Jijkoun, Maarten de Rijke

ISLA, University of Amsterdam

The task

- Given a target Wikipedia article, return important snippets about the target article from other Wikipedia articles of the same language or different languages
  - The snippets should be
    - *relevant*, *important*, *novel* with respect to the content of the target article, and *without duplicates*
- Main research aims/questions
  - Set up baselines
  - Compare techniques for determining relevant & important sentences

Monolingual WiQA

- System components

A. Identifying relevant sentences

B. Estimate sentence importance

C. Remove redundancy

A. Identifying relevant sentences

- Link-based method
  - Identify all sentences containing citations to the target
- Retrieval based method
  - Retrieve articles and take sentences containing the target title (string matching)
- Combination of the above
  - Retrieve articles
  - Sort articles based on retrieval score and take the top *n* articles
  - Take sentences containing citation to the target

B. Estimate sentence importance

- Combine m
  - Retrieval
  - Citation c
  - Sentence
  - Graph-ba

- Graph-based scoring
  - Assumptions
    - Sentence is important if it contains information typical for class of entities represented by target
    - What is typical (representative) is defined by a reference corpus
      - Random sample of Wikipedia articles falling under category labels of the target article
  - Computing score
    - Snippets in reference corpus vote for candidate snippets
    - Candidate snippets are ranked based on the number of votes they receive

C. Redundancy removal

- Assume snippets are sorted by importance score (Step B)
- A word-overlap score is computed between each candidate snippet and
  - snippets ranked above it, and
  - snippets in the target article
- Snippets with word-overlap score above a threshold value are discarded

Multilingual WiQA

- Monolingual runs on each language
- Multilingual similarity for redundancy removal
  - Generate bilingual lexicon
    - Corresponding page titles
    - Wikipedia re-direct feature used to identify synonyms
  - Compute cross-language similarity using the bilingual lexicon
  - Remove snippets above a certain threshold

Results

	Avg. yield	MRR	P@10
English			
Ret	2.938	0.523	0.329
Link	**3.385**	**0.579**	**0.358**
LinkRet	2.892	0.516	0.330
Dutch			
Ret	3.200	0.459	0.427
Link	**3.800**	**0.532**	**0.501**
LinkRet	3.500	0.532	0.494
English-Dutch			
LinkRet	5.03	0.518	0.535

Concluding remarks

- Decent baselines
- Techniques largely language independent
- Link-based method performed better
- Multilingual scores are higher than the monolingual scores
- Main sources of errors
  - Ambiguous titles (particularly for the retrieval approach)
  - Character encoding issues
  - Too few or too many initial candidates

Maarten de Rijke
mdr@science.uva.nl