

Towards Topic Driven Access to Full Text Documents

Caterina Caracciolo, Willem R. van Hage,
Maarten de Rijke

Informatics Institute
University of Amsterdam

September 15, 2004

Outline

- Motivations
- Strategy
- What kind of topic segmentation
- Annotation, segmentation, evaluation
- Results
- Discussion and future work

Motivation: Minimize Scrolling and Searching

Find a document, type Ctrl+F, press Enter, Enter, Enter...

Aims:

- “conceptual” retrieval,
- sensitive to reading needs

Applications:

- access long full text
- link ontologies to textual corpora

Strategy

Divide the text into coherent units (topic segmentation) and use them as a basis for retrieval.

- Today: topic segmentation in our case study:
 - *Handbook of Logic and Language* (1997. van Benthem, ter Meulen eds.)
 - manually annotated to mark topic shifts
 - *linear* paragraph aggregation: no overlap, no nesting
 - algorithms: TexTiling, C99

Linear Topic Segmentation Algorithms

- TextTiling (Hearst, ACL 1994)
 - repetition of words gives the structure of the document
 - vectorial similarity of blocks of text: looks at relative differences among similarity values
 - results in paragraph aggregation
- C99 (Choi, ANLP 2000)
 - matrix of similarity sentence by sentence (or paragraphs)
 - hierarchical divisive clustering

Annotation, Evaluation

- Two annotators annotated 2 chpt. (out of 20) resulting in a single annotation
- Loose guidelines: respect sections, mark paragraphs
- The numbers:
 - Chpt A: 168 paragraphs, divided into 102 segments;
 - Chpt B: 223 paragraphs, divided into 90 segments.
- Evaluation: Precision and Recall on segment breaks
 - quite crude and only quantitative
 - but well understood

Results

Chpt.	Baseline	TexTiling	C99
A	$P = .614$	$P = .602$	$P = .571$
	$R = 1$	$R = .803$	$R = .078$
	$F = .760$	$F = .683$	$F = .137$
B	$P = .408$	$P = .344$	$P = .565$
	$R = 1$	$R = .681$	$R = .142$
	$F = .579$	$F = .445$	$F = .228$

Baseline performs well with chapter A because segments are short. Chapter B is a difficult case: long paragraphs, narrative style.

Discussion

- Baseline scores well because of the annotation, C99 poorly
- In a quantitative evaluation on segment breaks, TexTiling gives better balance of P/R, with better results with more “formal” chapter. It has a clearer interpretation than C99
- Lack of user/reader oriented evaluation measure, i.e. related to readability and entry point (argumentation structure)

Conclusions and Future Work

- TextTiling seems to be the kind of algorithm to go for
- Writing style matters and should be addressed
- In progress: the collection of segments is used as a target for queries taken from the ontology
- Future: make use of structural information from the ontology

Pointer

LoLaLi project page accessible from:

www.science.uva.nl/~caterina/LoLaLi

Thanks!