Summarization

Abstract:

Automatic Summarization deals with the challenge of producing a shortened verison of a text or a collection of texts which the goal of including the most important information. While basic systems focus on identifying and extracting relevant sentences from the text, more advanced systems identify and abstract information from the source documents and attempt to produce shortened texts using natural language generation. The issue what constitutes important information is made explicit in so-called query- or question-based summarization systems, which attempt to summarize text under the perspective of a particular question.

This seminar is intended as an introduction to the active field of automatic summarization. We will explore the key issues and approaches and how they link to related topics such as question answering and paraphrasing. The question how to evaluation automatic summarization approaches will also play an important role.

Instructor: Detmar Meurers

- *Office:* Room 1.28, Blochbau (Wilhelmstr. 19)
- *Email:* dm@sfs.uni-tuebingen.de
- *Office hours:* Mondays 11:30–12:30

Course meets: in Seminarraum 1.13, Blochbau (Wilhelmstr. 19)
- Mondays, 16ct-18
- Wednesdays, 16ct-18

Credits and Campus:
- Credit Points: 10 (MA ISCL, term paper), 6 (BA ISCL, oral or written exam)

Syllabus (this file):

- html-Version (http://purl.org/dm/11/ss/sum)

- pdf-Version (http://purl.org/dm/11/ss/sum/syllabus.pdf)

Moodle page: https://moodle01.zdv.uni-tuebingen.de/course/view.php?id=592

Nature of course and our expectations: This is a Hauptseminar intended to provide an overview of the current approaches in this active research area. Each participant is expected to

1. regularly and actively participate in class, read the papers assigned by any of the presenters and post a question on Moodle to the "Reading Discussion Forum" on each reading *at the latest on the day before it is discussed* in class. (30% of grade for Hauptseminar, 50% for Proseminar)

> Note: Following the standard rules missing more than two meetings un-excused, automatically results in failing the class.

2. explore and present a topic (30% of grade for Hauptseminar, 50% for Proseminar):

   - select one of the sub-topics during the second week of the semester
   - research the topic, taking Detmar's references *as a starting point*
   - *a week before the presentation:*
     - prepare the presentation with slides and discuss the presentation with Detmar during his office hours
     - start a new Moodle thread on the "Reading Discussion Forum" specifying what every course participant should read to prepare for your presentation
   - present the topic in class

3. for a Hauptseminar Schein, work out a term paper (40% of grade for Hauptseminar)

   - select a topic and submit a one-page abstract *by Tuesday, July 12.*
     - The term paper topic must be related to the seminar topic.
   - email the term paper in pdf format to Detmar *before the beginning of the next semester, i.e., by September 30, 2011.*
     - Note for ISCL students: The term paper must be produced in LaTeX, and BibTex must be used for the bibliography.

Academic conduct and misconduct: Research is driven by discussion and free exchange of ideas, motivations, and perspectives. So you are encouraged to work in groups, discuss, and exchange ideas. At the same time, the foundation of the free exchange of ideas is that everyone is open about where they obtained which information. Concretely, this means you are expected to always make explicit when you've worked on something as a team – and keep in mind that being part of a team always means sharing the work.

For text you write, you always have to provide explicit references for any ideas or passages you reuse from somewhere else. Note that this includes text "found" on the web, where you should cite the url of the web site in case no more official publication is available.

Class etiquette: Please do not read or work on materials for other classes in our seminar. Come to class on time and do not pack up early. All portable electronic devices such as laptops and cell phones should be switched off for the entire length of the flight, oops, class. If for some reason, you must leave early or you have an important call coming in, or you have to miss class for an important reason, please let me know *before* class.

Topics

- Overview:

  Spärck Jones, Karen. "Automatic summarising: The state of the art." Information Processing and Management 43.6 (2007), 1449-1481.

- Keyphrase Extraction (Nico)

  - supervised
  - unsupervised

- Summarization approaches

  - Extraction-/Reduction-based summarization (Sarah Schulz)
  - Abstraction-based summarization (Michael)

- Query-/Question-based summarization (Aleks)

- Multi-document summarization (Agnia)

- Example systems (Dietmar):

  - MEAD
  - Newsblaster
  - GISTexter

- Evaluation of Summarization (Laura):

  - ROUGE
  - Pyramid Method

- Linguistic indicators for evaluating difficulty of input to summarize

- Paraphrase/Semantic similarity detection (Arseniy)

Sessions

Note that the following session plan is subject to change; it only constitutes the current state of our planning as the semester unfolds.

1. Wednesday, April 13

   - Syllabus and course motivation

2. Monday, April 18: Introduction [DETMAR]

   - Reading: Hovy (2005), http://en.wikipedia.org/wiki/Automatic_summarization

3. Wednesday, April 20: (cont.)

   Monday, April 25: Easter Monday Holiday (no class)

4. Wednesday, April 27: Overview based on Spärck Jones (Spyridoula Georgatou, Kaidi Loo)

5. Monday, May 2: (cont.)

6. Wednesday, May 4: (cont.)

7. Monday, May 9: Keyphrase extraction (Noko Schenk)

8. Wednesday, May 11: (cont.)

9. Monday, May 16: Comparing Meaning in Context (Ramon Ziai, Niels Ott)

10. Wednesday, May 18: (cont.)

11. Monday, May 23: Example systems (Dietmar Fiesel)

12. Wednesday, May 25: (cont.)

13. Monday, May 30: Extraction-/Reduction-based summarization (Sarah Schulz)

14. Wednesday, June 1: (cont.)

15. Monday, June 6: Absraction-based summarization (Michael Hahn, Laura Kassner)

16. Wednesday, June 8: (cont.)

    Monday, June 13: Pentacoste Hoilday (no class)

    Wednesday, June 15: Pentacoste Hoilday (no class)

17. Monday, June 20:

18. Wednesday, June 22:

19. Monday, June 27: Paraphrase detection (Arseniy Mstislavskiy)

20. Wednesday, June 29:

21. Monday, July 4: Query-based summarization (Aleksandar Dimitrov)

22. Wednesday, July 6: (cont.)

23. Monday, July 11: Evaluation of Summariation (Agnia Barsukova)

24. Wednesday, July 13:

25. Monday, July 18:

26. Wednesday, July 20:

References

Hovy, E. (2005). Automated Text Summarization. In R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford University Press, pp. 583–598.

Last update: May 25, 2011