Text Summarization and Information Access: Tools and Evaluation

Horacio Saggion - University of Sheffield - UK

- Motivation and Objectives:

  Recent years have witnessed an explosion of on-line unstructured information in multiple languages, making natural language processing technologies such as automatic text summarization increasingly important for the information society. Text Summarization provides users with condensed descriptions of documents, allowing them to make informed decisions based on text sumaries. Text summarization can be combined with Information Retrieval (IR) and Question Answering (QA) to provide users with focus-based or QA-based summaries which are targeted towards the users' specific needs. When the information a user looks for is spread across multiple sources, text summarization can be used to condense that information and present a non-redundant account of the most relevant facts found across a set of documents.

  The objective of this tutorial is to give an overview of a number of technologies in natural language processing for information access including: single and multi-document summarization, cross-lingual summarization; and summarization-based question answering.

  The tutorial will give an overview of summarization concepts and techniques as well as its relation and relevance to other technologies such as information retrieval and question answering. It will also include description of available resources for development, training and evaluation of summarization components. A summarization toolkit available from the lecturer's web site will be used for demonstration purposes. A number of summarization-based question answering components for the creation of definitional summaries and profiles summaries will also be demonstrated.

- Detailed content:

  - Introduction: information access technologies; natural language processing tools; information retrieval; information extraction; text summarization; question answering; NLP tools for information access: the Cubreporter Project as a case study in information access.
  - Summarization concepts: summary typology and examples; human factors in the production of summaries; human production and use.
  - Summarization techniques and systems: theoretical framework; superficial features: indicative phrases, term distribution, title, position, etc.; machine learning for sentence extraction; summarization by information extraction; text generation; language models; new techniques such as paraphrase identification and generation; cross-lingual summarization; summarization in languages other than English.

– Summarization resources: the SummBank corpus for cross-lingual summarization; the DUC corpus; available NLP tools to support summarization including a tool compatible with GATE.
– Evaluation methods: intrinsic and extrinsic methods; precision and recall; content-based metrics, the ROUGE package, the Pyramids evaluation.
– Evaluation exercises: the SUMMAC evaluation; the Document Understanding Conference (DUC) evaluation; the MSE 2005 evaluation; the Text Summarization Challenge; relation to TREC/QA evaluation.
– Summarization, Question Answering, and Information Retrieval: definitional question answering; profile-based summarization; TREC/QA evaluation; summarization for cross-document coreference.