Summarization and STS

Enrique Alfonseca
March 14, 2012

Automatic Text Summarization

Compressing textual information to show the data that is most important to the user:

- Extractive/Abstractive
- Single/multi-document.
- Headline vs. highlights vs. longer text.
- Query-focused.
- Opinion-focused.
- Update.
- Time-dependent.
- etc.

Automatic Text Summarization (II)

Highlighted components

- Relevance
- Redundancy

Both in extractive and non-extractive approaches.

(Lin and Bilmes, ACL-2011; many other componnents not mentioned here...)

In news MDS

- Relevance
  - Looking for "central" sentences, repeated/supported/similar to many other sentences in the collection.
  - Lots of redundancy.
  - Collections are mostly coherent.
  - Uncommon to see flat-out contradiction (but there are bias and updates).
- Redundancy
  - Are there similar pairs of sentences in the summary.
  - How similar is a candidate piece of information to what has already selected to be in the summary.

In extractive, news MDS

Trivial algorithm given a black-box STS:

- Define a sentence-pairwise similarity measure.
- Define a centrality measure (some "similarity mass" to other sentences in the document).
- Define an objective function for scoring summaries combining centrality + redundancy.
- Find the optimal summary (exponential search space: ad-hoc greedy, ILP-based or heuristic-based search).

- Also in supervised settings, e.g. learning a model optimizing a rouge score on a set of manual summaries.

Other scenarios

Sentence compression:

- Given a sentence, produce a sentence of smaller length in "category 4"
  - Possibly changing entity mentions with shorter nominal or pronominal mentions.
  - Removing constituents with unimportant details.
  - But keeping the sentences grammatical.

Sentence fusion:

- Similar scenario; t1 (or t2) spans 2 sentences.

But do we need STS for this?

- And, is grammaticality considered in STS at all?

Discussion

- Several summarization tasks can be formulated in terms of STS.
  - STS should be a good feature to have for summarization systems.
  - In several ways it is already used, maybe not under this name.
- More generally, underlying problems in solving RTE or STS are useful for summarization.