automatic summarization

di.unito.it
Dipartimento
di Informatica

credits

- E. Hovy, Chapter *Text Summarization*, in R. Mitkov (Ed.), The Oxford handbook of computational linguistics, Oxford University Press, 2005
- D. Jurafsky and J. H. Martin, *SPEECH and LANGUAGE PROCESSING, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2009
- Eduard Hovy and Daniel Marcu, ACL Tutorial on Text Summarization, ACL 1998, Université de Montréal Montréal, Québec, Canada

a definition

- The goal of text summarization is to produce an abridged version of a text which contains the important or relevant information.
  - an abstract of a scientific article, a summary of email threads, a headline for a news article, or the short snippets returned by web search engines to describe each retrieved document.

goals

- Indicative: give an idea of what is there, provides a reference function for selecting documents for more in-depth reading
- Informative: a substitute for the entire document, covers all the salient information in the source at some level of detail
- Critical: evaluates the subject matter of the source, expressing the abstractor's view on the quality of the work of the author

kinds of automatic summarization

- Extracts are summaries created by reusing portions (words, sentences, etc.) of the input text verbatim, while
- Abstracts are created by re-generating the extracted content
  - Paraphrase, generation

kinds of automatic summarization

- Output: User-focused (or topic-focused or query focused): summaries that are tailored to the requirements of a particular user or group of users
- Background: Does the reader have the needed prior knowledge?
  - Expert reader vs. Novice reader
- General: summaries aimed at a particular –usually broad – readership community

Summarisation approaches

- Shallow approaches
  - Syntactic level at most
  - Typically produce extracts
  - Extract salient parts of the source text and then arrange and present them in some effective manner
- Deeper approaches
  - Sentential semantic level
  - Produce abstracts and the synthesis phase involves natural language generation.
  - Knowledge-intensive, may require some domain specific coding

single doc versus multiple doc summarisation

- In single document summarisation we are given a single document and produce a summary.
  - Single document summarisation is thus used in situations like producing a headline or an outline, where the final goal is to characterise the content of a single document.
- In multiple document summarisation, the input is a group of documents, and our goal is to produce a condensation of the content of the entire group.
  - We might use multiple document summarisation when we are summarising a series of news stories on the same event, or whenever we have web content on the same topic that we'd like to synthesise and condense.

parameters

- Compression rate (summary length/source length)
- Audience (user-focused vs. generic)
- Relation to source (extract vs. abstract)
- Function (indicative vs. informative vs. critical)
- Coherence: the way the parts of the text gather together to form an integrated whole
  - Coherent vs. incoherent
  - Incoherent: unresolved anaphors, gaps in the reasoning, sentences which repeat the same or similar meaning (redundancy) a lack of organisation

approaches comparison

- NLP/IE:
  - Approach: try to 'understand' text—re-represent content using 'deeper' notation; then manipulate that.
  - Need: rules for text analysis and manipulation, at all levels.
  - Strengths: higher quality; supports abstracting.
  - Weaknesses: speed; still needs to scale up to robust open-domain summarisation.

- IR/Statistics:
  - Approach: operate at lexical level—use word frequency, collocation counts, etc.
  - Need: large amounts of text.
  - Strengths: robust; good for query-oriented summaries.
  - Weaknesses: lower quality; inability to manipulate information at abstract levels.

relevance criteria

Position in the text

- Important sentences occur in specific positions
  - *"lead-based" summary* (just take first sentence(s)!)
  - Important information occurs in specific sections of the document (introduction/conclusion)
  - Experiments:
    - In 85% of 200 individual paragraphs the topic sentences occurred in initial position and in 7% in final position

Title method

- Title of document indicates its content
  - Not true for novels usually
  - What about blogs …?
- Words in title help find relevant content
  - Create a list of title words, remove "stop words"
  - Use those as keywords in order to find important sentences

Optimum Position Policy (OPP)

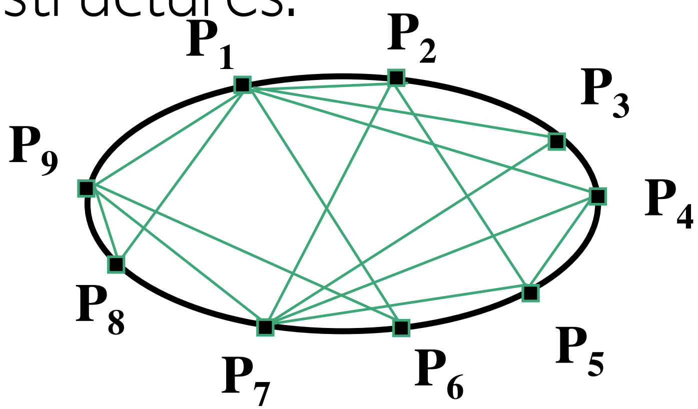
- Relevant sentences are located at positions that are genre-dependent; these positions can be either known or determined automatically through training
  - Step 1: For each article, determine the overlap between sentences and the index terms (e.g., title terms)
  - Step 2: Determine a partial ordering over the locations where sentences containing important words occur: Optimal Position Policy (OPP)

Cue phrases method

- Important sentences contain cue words/indicative phrases,
  - "The main aim of the present paper is to describe…"
  - "The purpose of this article is to review…"
  - "In this report, we outline…"
  - "Our investigation has shown that…"
- Some words are considered bonus others stigma
  - bonus: comparatives, superlatives, conclusive expressions, etc.
  - stigma: negatives, pronouns, *etc.* non-important sentences contain 'stigma phrases' such as hardly and impossible.
- These phrases can be detected automatically
- Method: Add to sentence score if it contains a bonus phrase, penalise if it contains a stigma phrase.

Cohesion-based methods

- Important sentences/paragraphs are the highest connected entities in more or less elaborate semantic structures.
- Classes of approaches
  - word co-occurrences;
  - local salience and grammatical relations;
  - co-reference;
  - lexical similarity (WordNet, lexical chains);
  - combinations of the above.



Cohesion: word co-occurrence

- Apply IR methods at the document level: texts are collections of paragraphs
  - Use a traditional, IR-based, word similarity measure to determine for each paragraph P_i the set S_i of paragraphs that P_i is related to.
- Method:
  - determine relatedness score S_i for each paragraph,
  - extract paragraphs with largest S_i scores.

3 (to 1) steps

- Text summarisation systems are generally described by their solutions to the following three problems:
- *Content Selection*: What information to select from the document(s) we are summarising. We usually make the simplifying assumption that the granularity of extraction is the sentence or clause. Content selection thus mainly consists of choosing which sentences or clauses to extract into the summary.
- *Information Ordering*: How to order and structure the extracted units.
- *Sentence Realisation*: What kind of clean up to perform on the extracted units so they are fluent in their new context.

unsupervised algorithm

- The simplest unsupervised algorithm is to select sentences that have more salient or informative words.
  - Sentences that contain more informative words tend to be more extract-worthy.
- *Saliency* is usually defined by computing the topic signature, a set of salient or signature terms, each of whose saliency scores is greater than some threshold θ.
  - Saliency could be measured in terms of simple word frequency, but frequency has the problem that a word might have a high probability in English in general but not be particularly topical to a particular document.
- Lexical specificity can thus be adopted in order to individuate the most salient terms, and to score the sentences where they appear.

a simple *extractive* algorithm

- reduce the document size of e.g., 10%, 20%, 30%

1. individuate the topic of the text being summarised; the topic can be referred to as a (set of) NASARI vector(s):

   v_{t1} = {*term*$_1$*_score*, *term*$_2$*_score*, …, *term*$_{10}$*_score* }
   v_{t2} = {*term*$_1$*_score*, *term*$_2$*_score*, …, *term*$_{10}$*_score* }
   …

2. create the context, by collecting the vectors of terms herein (this step can be repeated, by dumping the contribution of the associated terms at each round);

3. retain paragraphs whose sentences contain the most salient terms, based on the Weighted Overlap, $WO(v_1,v_2)$
   - rerank paragraphs weight by applying at least one of the mentioned approaches (*title*, *cue*, *phrase*, *cohesion*).

NASARI (lexical) subset

- two distribution files are provided for NASARI, that require different resources allocation.
  - *dd-nasari.txt*. a subset of NASARI (obtained by truncating vectors at 10 features). 3,587,754 vectors, ~600MB;

    https://goo.gl/85BubW
  - *dd-small-nasari-15.txt*. a subset of NASARI. same filtering as above, with 15 features + intersection with 60K lemmas in the Corpus of Contemporary American English: 13,084 vectors, 2MB storage (many entities removed here…).
- the second one has been extracted for starting our experimentation; the second one is intended to explore the resource in a richer (though reduced) flavour.

documents for summarisation

- text documents are provided for summarisation purposes:
  - *Andy-Warhol.txt*
  - *Ebola-virus-disease.txt*
  - *Life-indoors.txt*
  - *Napoleon-wiki.txt*
  - *Trump-wall.txt*
- do experiment with different compression rates: 10%, 20% and 30%.

evaluation

- evaluation can be performed based on two complimentary metrics
  - BLEU (bilingual evaluation understudy) regarding precision; and
  - ROUGE (Recall-Oriented Understudy for Gisting Evaluation) as regards as recall.

BLUE (bilingual evaluation understudy)

- scoring function that has been worked out to assess systems for automatic translation
  - build a reference summary, as a list of relevant terms that should be present.
  - compare the set of terms in the automatic summary (which we call candidate summary,) to those in the candidate summary.
  - the BLEU score is computed as $P = m/w_t$ that is the fraction of terms from the candidate that are found in the reference, where m is the number of terms in the candidate that are in the reference, and w_t is the size of the candidate
- precision in IR is customarily defined as

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

- This metrics estimates in how far the words (and/or n-grams) in the human reference summaries appeared in the summaries built by the system
  - ROUGE-N: Overlap of N-grams between candidate and reference summary.
  - ROUGE-1 refers to the overlap of unigram (each word) between the system and reference summaries.
- recall in IR is customarily defined as

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$