An Abstract Argumentation-based Approach to Automatic Extractive Text Summarization

Stefano Ferilli, Andrea Pazienza
Dipartimento di Informatica - Università di Bari, Bari, Italy

stefano.ferilli@uniba.it

Overview

- Introduction & Motivation
- Proposed solution
  - Abstract Argumentation
- Evaluation
- Conclusions & Future Work

Introduction

- Text summarization
  - The process of automatically creating a shorter version of one or more text documents
- 2 approaches
  - Extractive: work by selecting sentences from the input document(s) according to some criterion
  - Abstractive: may produce summaries containing sentences that were not present in the input document(s)

Motivation

- Information overload problem
  - Legacy and New documents
    - Internet 2.0
  - News
- Reading/Analyzing/Understanding the available documents impossible for humans
  - Huge quantity
  - Fast rate

Objectives

- Generality
  - Language-independent approach
- Automatic approach
  - Summary length
- Desirable properties
  - Diversity: each sentence in the summary should bring additional information
  - Coverage: the sentences in the summary should contain all the relevant information from the original text

Proposed Solution

- Abstract Argumentation
  - A non-monotonic inferential strategy aimed at selecting reliable items in a set of conflicting claims
- Idea
  - Sentences ~ Arguments
  - Conflicts ~ Sentences that should not be both included in the same summary
    - Criterion: similarity
      - Attacks between pairs of similar sentences (to enforce *diversity*)
      - Supports between pairs of dissimilar sentences (to enforce *coverage*)

Abstract Argumentation Theory

- Argumentation Frameworks (AF) ~ Graphs
  - Arguments ~ Nodes, Attacks ~ Edges
  - Bipolar AFs: Consider both attacks and supports
  - Weighted AFs: Strength of attacks (or supports)
  - Semantics: compute subsets of arguments (Extensions) that are mutually compatible
    - Several options, more skeptic or more credulous
- BAFs the simplest AF that allows to consider both attacks and supports between arguments

Proposed Approach

- NLP Pipeline
  - Sentence splitting
  - Tokenization
  - Lemmatization
  - Stopword removal
  - Word Sense Disambiguation
- Argumentation Framework Building & Evaluation

Proposed Approach

- Computation of similarity between pairs of sentences
  - Based on similarity of their building tokens (words)
  - Similarity between tokens computed as a linear combination of 3 similarity functions:
    - Syntactic
      - based on the Jaccard Index applied to syntactic dependencies
    - Semantic
      - based on taxonomic information using synsets in WordNet
    - Embedding
      - based on the cosine similarity between word embeddings
  - Normalized into [0,1]

Proposed Approach

- Argumentation Framework building
  - Heuristic inspired by the concept of *inconsistency budget* in WAFs
    - 2 thresholds α,β with $\beta < \alpha$
      - Attack threshold $\alpha \in [0, 1]$
        - Edges with weight $\geq \alpha$ ~ Attacks
      - Support threshold $\beta \in [0, 1]$
        - Edges with weight $\leq \beta$ ~ Supports
    - Intermediate similarity range between β and α does not generate attacks nor supports

Proposed Approach

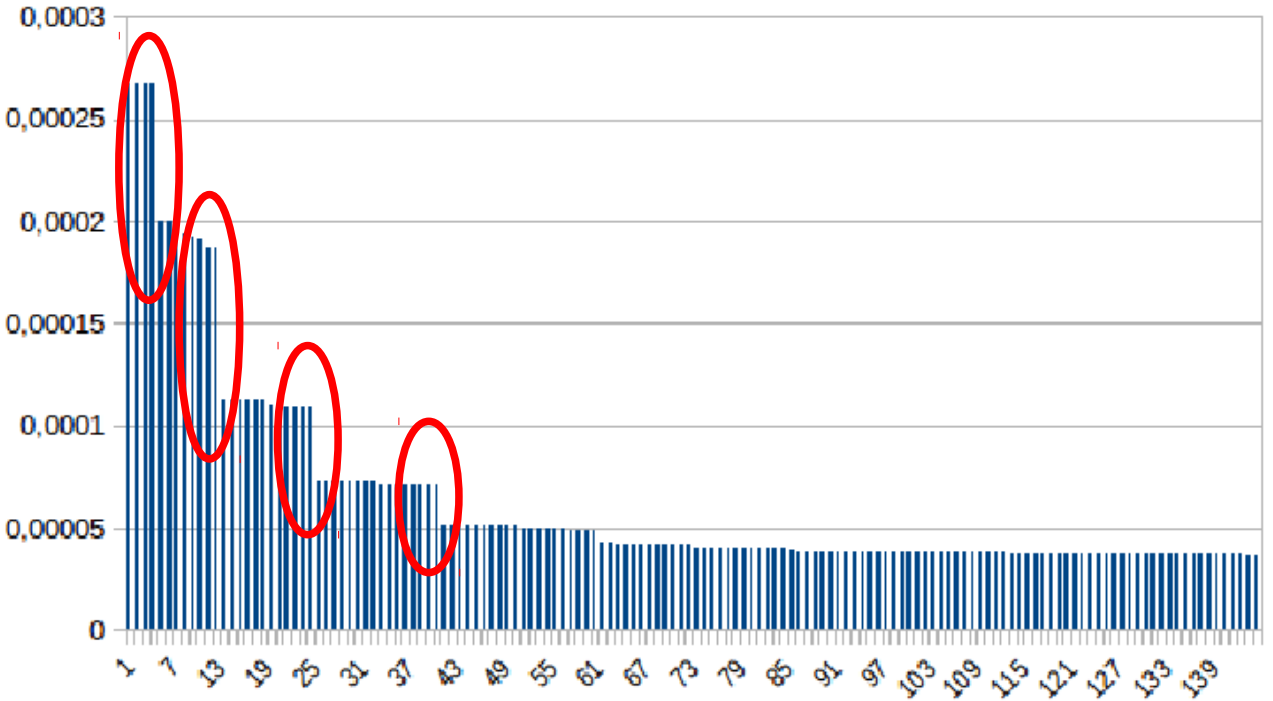
- Semantics Evaluation
  - Conflict-free :
    - Rewards sentences that maximize diversity
  - d-/s-admissible :
    - May exceed the allowed length
  - Complete :
    - May include sets of arguments that are too small
  - d-/s-preferred :
    - Should improve quality and length wrt admissible and complete
  - Stable :
    - Should select the most dissimilar sentences
    - Might not achieve any solution at all

Evaluation

- Dataset
  - Single-document text summarization task of the English version of the MultiLing 2015 dataset
- Performance Measures
  - ROUGE-*n*
    - Percentage of overlapping word *n*-grams between generated summary and ground truth
  - Quality(*s*) = ROUGE-1(*s*) / length(*s*)
    - Compound indicator defined to penalize long summaries, since our approach automatically determines summary length

Evaluation

- Quality results for 144 argumentation settings
  - **X** = summarization settings (σ,α,β)
    - $\sigma \in$ {s-admissible, d-admissible, stable, complete}
    - $\beta \in [0.1,0.8]$, $\alpha \in [0.2,0.9]$, step 0.1, $\beta < \alpha$
  - **Y** = average quality on all settings run for that task
    - by decreasing value
      - 4 steps



Evaluation

- Results on selected settings (steps)
  - Average size of full texts: 25542 characters
  - Average size of the ground truths: 1857
    - 7% of the full texts

	ROUGE-1	ROUGE-2
WORST	37.17%	9.93%
BEST	50.38%	15.10%
ORACLE	61.91%	22.42%

Step	Semantics	α	β	length (%)	Quality	Rouge-1		Rouge-2	
						Recall	Precision	Recall	Precision
1	s-admissible d-admissible	0.1	0.3	1279 (5%)	2.00E-04	25.57%	41.97%		
2	s-admissible d-admissible	0.1	0.4	4365 (17%)	1.13E-04	49.28%	30.32%	15.49%	7.22%
3	stable complete	0.1	0.5	8544 (33%)	7.07E-05	60.44%	24.44%	23.98%	7.43%
4	s-admissible d-admissible	0.1	0.5	9826 (38%)	7.33E-05	72.09%	26.65%	27.26%	6.16%

Evaluation

- 4 steps in the graphic
  - Sudden drop in quality
    - Interesting for further investigation
  - Each corresponds to 2 different semantics that returned exactly the same results
  - No summary with length close to that of the ground truth
    - First step are shorter
    - Second step length jumps from 1352 to 4365 characters

Evaluation

- Proposed approach sensible and effective in returning relevant summaries, and competitive in performance, albeit paying in summary length
  - ROUGE-1
    - Step 1: recall not bad, precision quite high, even if the length of the summary is less than that of the ground truth
    - Step 2: comparable to the state-of-the-art, but using 17% (more than twice the ground truth)
    - Step 3: comparable to Oracle, but using the 33% (1/3) of the text
    - Step 4: much (> 10%) better than Oracle, but using 38% (less than 2/5) of the input texts
      - Able to catch nearly 3/4 of the content
  - ROUGE-2:
    - Same as above, plus
    - Recall slightly larger than the reference systems

Evaluation

- Disclaimer
  - Ground truth in this dataset obtained by humans using an *abstractive* approach
    - Extractive text summarization procedure possibly inappropriate
      - No exact match between the sentences in the input text and those in the summary
      - Summaries may contain words that are not contained in the input text
      - Summarization process extremely subjective
        - Many summaries may be appropriate for a given input text
        - Only one provided as ground truth

Evaluation

- Semantics
  - s-preferred and d-preferred provide relevant results
    - Confirms our hypothesis
  - stable and complete may also yield interesting results
    - Trade-off corresponding to the performance of ORACLE
- Qualitative evaluation: humans reported that
  - the proposed summaries have little redundancy,
  - yet provide a sensible account of the original document,
  - also ensuring smooth discourse flow,
    - even if obtained by filtering out sentences that, since present in the original text, presumably included relevant parts as regards the content and/or the flow of discourse

Conclusions

- Problem: Huge and ever-increasing number of documents in Digital Libraries
  - Impossible for humans to read them in order to assess their relevance and/or grasp their content
  - Solution: Automatic Text Summarization
    - Proposal: extractive approach based on Abstract Argumentation, and on similarity between sentences
- Results: viable and effective
  - Autonomously determines the number of sentences to be included in the summary
    - Summary typically larger than the dataset's ground truth, but still significantly shorter than the original text

Future Work

- Exploring other argumentation frameworks (e.g., those that may handle weights on attacks and supports) and semantics
- Further filtering of results returned by the argumentation-based selection
- Running experiments on other datasets and languages