The Role of Semantics in Automatic Summarization: A Feasibility Study

Tal Baumel
Computer Science Department
Ben-Gurion University
Beer-Sheva, Israel
talbau@cs.bgu.ac.il

Michael Elhadad
Computer Science Department
Ben-Gurion University
Beer-Sheva, Israel
elhadad@cs.bgu.ac.il

Abstract

State-of-the-art methods in automatic summarization rely almost exclusively on extracting salient sentences from input texts. Such extractive methods succeed in producing summaries which capture salient information but fail to produce fluent and coherent summaries. Recent progress in robust semantic analysis makes the application of semantic techniques to summarization relevant. We review in this paper areas in the field of summarization that can benefit from the introduction of the type of semantic analysis that has become available. The main pain points that semantic information can alleviate are: aiming for more fluent summaries by exploiting logical form representation of the source text; identifying salient information and avoiding redundancy by relying on textual entailment and paraphrase identification; and generating a coherent summary while relying on rhetorical structure and discourse structure information extracted from the source documents. In addition, we review the possibility to perform automatic Pyramid evaluation of summarization quality that relies on robust semantic similarity measures.

Introduction

Current state-of-the-art automatic multi-document summarization methods (Erkan and Radev, 2004; Haghighi and Vanderwende, 2009) and automatic evaluation methods (Lin, 2004) rely exclusively on lexical features. Those methods cannot attend the following problems:

- **Non-lexical similarity cannot be recognized:** pairs such as "*Danny and his friends bought a chocolate bar from Yossi*" and "*He sold them a snack*" are quasi-paraphrases although they don't share any lexical words. Correct SRL annotation can identify that *"buy"* and "*sell*" refer to the same frame and appropriate pronoun resolution would match the pronouns to the correct entities;
- **High lexical coverage does not mean coherent summaries:** the focus on current summarization research has been to optimize ROUGE results. ROUGE measures lexical overlap between manual summaries and the candidate one. We observe, however, that summarizers that achieve the highest ROUGE scores often obtain low manual scores because they produce non-cohesive and sometimes non-coherent summaries.

We address those limitations of existing methods based exclusively on lexical analysis by using intermediate semantic annotations for both summarization generation and evaluation.

Summarization Generation

For generating automatic summarizations, we propose the following scheme:

- **Annotate the input documents with SRL**: we annotated large summarization datasets (DUC and TAC) using SEMAFOR (Das *et al.*, 2014) for FrameNet annotations and SENNA (Collobert *et al.*, 2011) for PropBank annotations. We parse both the source documents and the manual summaries of each document cluster. We are producing a descriptive analysis of the distribution of Frames and Entities in the datasets.
- **Apply salience identifier:** both PropBank and FrameNet annotations help with identifying paraphrases - syntactic alternations or lexical variants of the same relations (*i.e.*, "add up" and "total" are both different lexical units of the frame "Adding_up"). FrameNet annotations include frames relations such as: Inherits from, Perspective on,

Uses that can be further exploited by our system to align predicates across documents and with manual summaries.

- **Generate summarization given salient frames:** after identifying the salient frames we generate coherent text using text generations tools such as FUF (Elhadad, 1991)

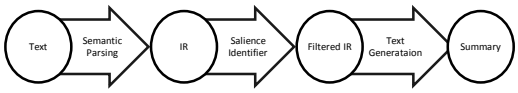

Figure 1: Proposed Summarization Scheme

Summarization Evaluation

For automatic summarization evaluation, we adapt the MEANT approach developed in Machine Translation (Wu, 2011). MEANT measures translation quality as follows: First, semantic role labeling is performed (either manually or automatically) on both reference translation and machine translation. The annotated translations are compared by matching frame by frame, argument by argument. The final score is the F-measure score of the aligned graph. When using manual SRL, MEANT can achieve 0.4324 Kendall $\boldsymbol{\tau}$ correlation agreement with human evaluation, same as HTER. When using an automatic SRL tool (Pradhan et al., 2004), it achieves 0.3423 agreement with manual annotators, while BLEU only achieves 0.1982. Adapting a similar system to measure the semantic quality and coherence of automatic summary requires modification in the alignment procedure and weighting predicates according to the number of times they appear in manual summaries (in the manner similar to the Pyramid evaluation method (Nenkova *et al.*, 2007), but taking into account the confidence in the semantic match across content units.

While MEANT is an important tool for text similarity, it has only been tested on automatic translation datasets, those datasets tested translations of single sentences and not 250 word text as is common in automatic summarization datasets. There are new challenges that must be addressed when evaluating large texts:

- **Referring expression:** natural text usually contains anaphors (*e.g.*, "The boy went to school. He had fun there"). Anaphors are frequent and must be resolved to match content units in the text and to verify that the generated text does not introduce unwanted ambiguity.
- **Document structure:** Knowing what information should be included in the summary is not sufficient to construct coherent text. In order to convey a narrative in the summary we need to determine which information should appear where, and what fragments need to be in the same paragraphs.
- **Aggregation:** merging similar fragments across frames improves readability and avoids redundancy. Aggregation appears at various levels of language: lexical, syntactic and in discourse structure (*i.e.*, "Jacob is the son of Isaac" and "Esau is the son of Isaac" aggregate into "Jacob and Esau are sons of Isaac"). The evaluation system should recognize that frame-elements from different sentences are aggregated into a single proposition.

We are measuring the frequency of each of these points on the automatically SRL annotated datasets of DUC and TAC. On the basis of this data analysis, we are assessing the potential of developing better automatic evaluation of summarization algorithms, and to improve each of the candidate aspects of the process.

References

[1] Collobert, Ronan, et al. "Natural language processing (almost) from scratch." The Journal of Machine Learning Research 12 (2011): 2493-2537. Baron, Jason R., David D. Lewis, and Douglas W. Oard. "TREC 2006 Legal Track Overview." TREC. 2006.

[2] Das, Dipanjan, et al. "Frame-semantic parsing." Computational Linguistics 40.1 (2014): 9-56.

[3] Elhadad, Michael. "FUF: The universal unifier user manual version 5.0." (1991).

[4] Erkan, Günes, and Dragomir R. Radev. "LexRank: graph-based lexical centrality as salience in text summarization." Journal of Artificial Intelligence Research (2004): 457-479.

[5] Haghighi, Aria, and Lucy Vanderwende. "Exploring content models for multi-document summarization." Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.

[6] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. 2004.

[7] Lo, Chi-kiu, and Dekai Wu. "MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Tech-

nologies-Volume 1. Association for Computational Linguistics, 2011.

[8] Ani Nenkova, Rebecca Passonneau and Kathy McKeown. The Pyramid Method: Incorporating human content selection variation in summarization evaluation, Journal ACM Transactions on Speech and Language Processing (TSLP) Volume 4 Issue 2, May 2007

[9] Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow Semantic Parsing Using Support Vector Machines. In Proceedings of the 2004 Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04), 2004.