Building Conceptual Maps from Scientific Articles

Jumana Nassour-Kassis and **Michael Elhadad**
Dept. of Computer Science
Ben-Gurion University
Beer-Sheva, Israel
{jumanan,elhadad}@cs.bgu.ac.il

Arnon Sturm
Dept. of Information Systems Engineering
Ben-Gurion University
Beer-Sheva, Israel
sturm@bgu.ac.il

Abstract

We present a new form of summarization for scientific domains in the form of conceptual maps which model the hierarchy and the relations among problems and solutions in a domain as presented in scientific articles.

1 Introduction

Scientific articles are a very useful source of information. Articles are written in a "semi-structured" manner, that is, they follow a standardized rhetorical structure (abstract, introduction, motivation, related work, citations), but are presented in free text. The high variability of the form of scientific articles complicates the task of automatically extracting information from them.

We adopt a general purpose ontological schema describing problems, solutions and their inter-relation. We attempt to generate a conceptual map of a given domain using information extracted from a collection of articles annotated with the proposed schema, which offers a visual summarization of the domain. Figure 1 shows an example of such a conceptual map created manually from ten annotated abstracts in the field of summarization.

2 Related Work

In the field of automatic analysis of scientific articles, some approaches have focused on understanding scientific argumentation, which extracts sentences belonging to a certain rhetorical zone, but does not attempt to model what the sentences say (Gupta and Manning, 2011; Teufel and Moens, 1999). In our work, we exploit the rhetorical structure of articles, but also attempt to model the hierarchy and the relations among the problems and solutions proposed in a specific scientific domain.

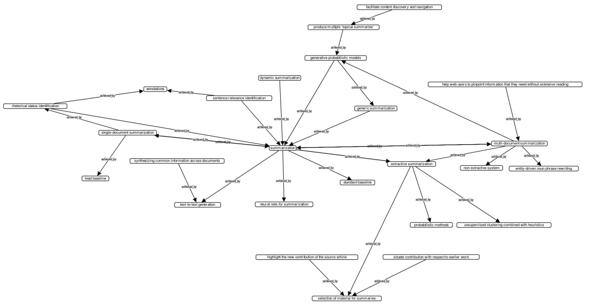
facilitate content discovery and navigation
achieved_by
produce multiple 'topical summaries'
achieved_by
generative probabilistic models
dynamic summarization
help web users to pinpoint information that they need without extensive reading
annotations
achieved_by
achieved_by
sentence relevance identification
rhetorical status identification
achieved_by
generic summarization
achieved_by
achieved_by
achieved_by
achieved_by
achieved_by
single-document summarization
achieved_by
achieved_by
achieved_by
summarization
achieved_by
achieved_by
multi-document summarization
achieved_by
achieved_by
extractive summarization
achieved_by
achieved_by
achieved_by
non-extractive system
entity-driven noun phrase rewriting
synthesizing common information across documents
achieved_by
lead baseline
achieved_by
achieved_by
standard baseline
achieved_by
text-to-text generation
neural nets for summarization
achieved_by
achieved_by
probabilistic methods
unsupervised clustering combined with heuristics
achieved_by
highlight the new contribution of the source article
situate contribution with respect to earlier work
achieved_by
achieved_by
selection of material for summaries

Figure 1: A map of 10 abstracts about summarization, created by merging 10 different maps with some clustering and pruning.

Other approaches focus on deep semantic models of a specific domain and attempt to perform full semantic interpretation of text to construct a formal model of this domain (Berant et al., 2014; Scaria et al., 2013). We intend to cover a variety of unrelated domains. As a result, we do not aim to construct a full detailed semantic model of the domain - capable of sophisticated inference and temporal reasoning. Instead, we aim to capture systematic relations across problems and solutions.

The construction of the proposed conceptual maps is closely related to the field of multi-document automatic text summarization (Qazvinian et al., 2013; Jha et al., 2015) - as it aims to identify central information units within a collection of source textual documents, and to present this information in a coherent manner. In contrast to summarization though, we propose to present the central information in a formal ontology and a visual representation as opposed to a textual manner.

3 Annotation Schema

We adopted a simple domain-independent annotation schema which includes six types of relations, and two types of entities: tasks or attributes. Tasks represent problems, *e.g.*, *summarization*, or solutions, *e.g.*, *extractive summarization*. Attributes are properties of the tasks, *e.g.*, *effectiveness*, *indicativeness*, *rouge* (the name of a quality measure for text summarization). The relations among entities are: Means-End, Instance-of, Consists-of, Associated-with, Contributes-to, and Compares-to. Table 2 illustrates annotations extracted from a sample text sample.

Index	Sentence
0	This paper analyzes the topic identification stage of single-document automatic text summarization across four different domains, consisting of newswire, literary, scientific and legal documents.
1	We present a study that explores the summary space of each domain via an exhaustive search strategy, and finds the probability density function (pdf) of the ROUGE score distributions for each domain.
2	We then use this pdf to calculate the percentile rank of extractive summarization systems.
3	Our results introduce a new way to judge the success of automatic summarization systems and bring quantified explanations to questions such as why it was so hard for the systems to date to have a statistically significant improvement over the lead baseline in the news domain.

Table 1: Sentences of an abstract of one of the annotated articles, in the field of summarization.

4 Methods and Datasets

To test the reliability of the scheme, two annotators manually annotated abstracts of twenty scientific articles in the field of summarization. After a few iterations, where we adapted the schema and guidelines, the annotators reached high agreement, and the extracted information proved to be informative. We are in the process of annotating another dataset consisting of ten articles in the field of cyber-security.

Sentence	Relation	Arguments
0	MEANS-END	- **Target**: summarization - **Means**: single document summarization
0	CONSISTS-OF	- **Parent**: single-document automatic text summarization - **Subtasks**: topic identification - **Context**: four different domains consisting of newswire literary scientific and legal documents
1	INSTANCE-OF	- **Instance**: x - **Type**: a study that explores the summary space of each domain
1	MEANS-END	- **Target**: x - **Means**: an exhaustive search strategy
1	MEANS-END	- **Target**: find the probability density function (pdf) of the ROUGE score distributions - **Means**: x
2	MEANS-END	- **Target**: calculate the percentile rank of extractive summarization systems - **Means**: find the probability density function (pdf) of the ROUGE score distributions
2	MEANS-END	- **Target**: summarization - **Means**: extractive summarization
3	MEANS-END	- **Target**: x - **Means**: judge the success of automatic summarization systems
3	MEANS-END	- **Target**: single-document summarization - **Means**: lead baseline,news domain
3	ASSOCIATED-WITH	- **Target**: summarization - **Property**: quality
3	COMPARES-TO	- **Tasks1**: summarization systems - **Task2**: lead baseline - **Property** quality - **Associated-with**: summarization - **Comparison-Value**: < - **Context**: news domain

Table 2: Some of the relations extracted, using our annotation scheme.

Automatic annotation according to the adopted schema requires a variety of NLP techniques: shallow semantic parsing to extract predicates and their arguments; entity linking to align extracted entities to a repository of candidate entities in the domain; discourse parsing to extract relations, mainly Means-End, across semantic predicates; terminology extraction to identify task names; and topic-based text segmentation to facilitate relation extraction. We have experimented with state of the art tools to achieve these tasks and are in the process of consolidating their output into the target annotations described above.

Given an annotated collection of articles on a given domain, we aim to generate a conceptual map summarizing that domain. Simply generating a map containing all the relations described in every article, proved to give unreadable results. Hence, the main challenge lies in deciding which relations to keep, and how to aggregate entities and relations to create a readable conceptual map that briefly covers the key aspects of the domain. We are adapting multi-document summarization techniques to this task and designing evaluation metrics to assess its effectiveness.

References

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Brad Huang, Christopher D Manning, Abby Vander Linden, Brittany Harding, and Peter Clark. 2014. Modeling biological processes for reading comprehension. In *Proc. EMNLP*.

Sonal Gupta and Christopher D Manning. 2011. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *IJCNLP*, pages 1–9.

Rahul Jha, Reed Coke, and Dragomir Radev. 2015. Surveyor: A system for generating coherent survey articles for scientific topics. *Ann Arbor*, 1001:48109.

Vahed Qazvinian, Dragomir R Radev, Saif Mohammad, Bonnie J Dorr, David M Zajic, Michael Whidby, and Taesun Moon. 2013. Generating extractive summaries of scientific paradigms. *J. Artif. Intell. Res.(JAIR)*, 46:165–201.

Aju Thalappillil Scaria, Jonathan Berant, Mengqiu Wang, Christopher D Manning, Justin Lewis, Brittany Harding, and Peter Clark. 2013. Learning biological processes with global constraints. In *Proceedings of EMNLP*.

Simone Teufel and Marc Moens. 1999. Discourse-level argumentation in scientific articles: human and automatic annotation. In *In Proceedings of the ACL-1999 Workshop Towards Standards and Tools for Discourse Tagging*. Citeseer.