Extractive Summaries for Educational Science Content

Sebastian de la Chica, Faisal Ahmad, James H. Martin, Tamara Sumner
Institute of Cognitive Science
Department of Computer Science
University of Colorado at Boulder
sebastian.delachica, faisal.ahmad, james.martin,
tamara.sumner@colorado.edu

Abstract

This paper describes an extractive summarizer for educational science content called COGENT. COGENT extends MEAD based on strategies elicited from an empirical study with domain and instructional experts. COGENT implements a hybrid approach integrating both domain independent sentence scoring features and domain-aware features. Initial evaluation results indicate that COGENT outperforms existing summarizers and generates summaries that closely resemble those generated by human experts.

1 Introduction

Knowledge maps consist of nodes containing rich concept descriptions interconnected using a limited set of relationship types (Holley and Dansereau, 1984). Learning research indicates that knowledge maps may be useful for learners to understand the macro-level structure of an information space (O'Donnell et al., 2002). Knowledge maps have also emerged as an effective computational infrastructure to support the automated generation of conceptual browsers. Such conceptual browsers appear to allow students to focus on the science content of large educational digital libraries (Sumner et al., 2003), such as the Digital Library for Earth System Education (DLESE.org). Knowledge maps have also shown promise as domain and student knowledge representations to support personalized learning interactions (de la Chica et al., 2008).

In this paper we describe our progress towards the generation of science concept inventories as summaries of digital library collections. Such inventories provide the basis for the construction of knowledge maps useful both as computational knowledge representations and as learning resources for presentation to the student.

2 Related Work

Our work is informed by efforts to automate the acquisition of ontology concepts from text. OntoLearn extracts candidate domain terms from texts using a syntactic parse and updates an existing ontology with the identified concepts and relationships (Navigli and Velardi, 2004). Knowledge Puzzle focuses on n-gram identification to produce a list of candidate terms pruned using information extraction techniques to derive the ontology (Zouaq et al., 2007). Lin and Pantel (2002) discover concepts using clustering by committee to group terms into conceptually related clusters. These approaches produce ontologies of very fine granularity and therefore graphs that may not be suitable for presentation to a student.

Multi-document summarization (MDS) research also informs our work. XDoX analyzes large document sets to extract important themes using n-gram scoring and clustering (Hardy et al., 2002). Topic representation and topic themes have also served as the basis for the exploration of promising MDS techniques (Harabagiu and Lacatusu, 2005). Finally, MEAD is a widely used MDS and evaluation platform (Radev et al., 2000). While all these systems have produced promising results in automated evaluations, none have directly targeted educational content collections.

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 17–20,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

3 Empirical Study

We have conducted a study to capture how human experts processed digital library resources to create a domain knowledge map. Four geology and instructional design experts selected 20 resources from DLESE to construct a knowledge map on earthquakes and plates tectonics for high school age learners. The resulting knowledge map consists of 564 concepts and 578 relationships.

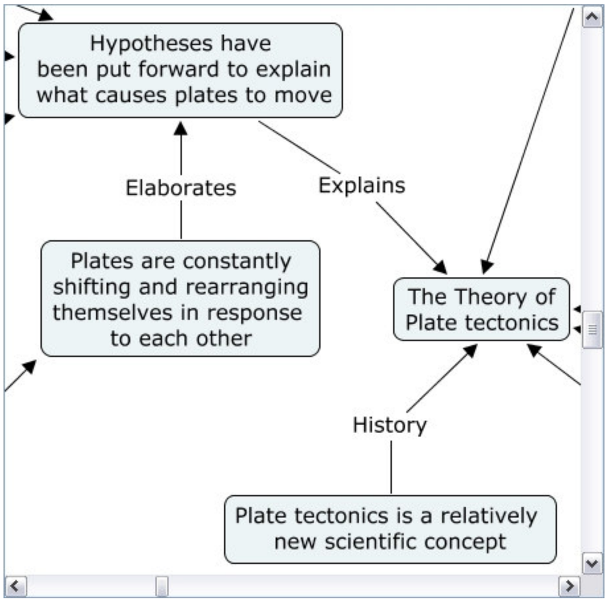
Hypotheses have been put forward to explain what causes plates to move
Elaborates
Explains
Plates are constantly shifting and rearranging themselves in response to each other
The Theory of Plate tectonics
History
Plate tectonics is a relatively new scientific concept

Figure 1. Expert knowledge map excerpt

The concepts include 7,846 words, or 5% of the resources. Our experts relied on copying-and-pasting (58%) and paraphrasing (37%) to create most concepts. Only 5% of the concepts could not be traced directly to the original resources. Relationship types were used in a Zipf-like distribution with the top 2 relationship types each accounting for more than 10% of all relationships: elaborations (19%) and examples (14%).

Analysis by an independent instructional expert indicates that this knowledge map provides adequate coverage of nationally-recognized educational goals on earthquakes and plate tectonics for high school learners using the American Association for the Advancement of Science (AAAS) Benchmarks (Project 2061, 1993).

Verbal protocol analysis shows that all experts used external sources to create the knowledge map, including their own expertise, other digital library resources, and the National Science Education Standards (NSES), a comprehensive collection of nationally-recognized science learning goals for K-12 students (National Research Council, 1996).

We have examined sentence extraction agreement between experts using the prevalence-adjusted bias-adjusted (PABA) kappa to account for prevalence of judgments and conflicting biases amongst experts (Byrt et al., 1993). The average PABA-kappa value of 0.62 indicates that experts substantially agree on sentence extraction from digital library resources. This level of agreement suggests that these concepts may serve as the reference summary to evaluate our system.

4 Summarizer for Science Education

We have implemented an extractive summarizer for educational science content, COGENT, based on MEAD version 3.11 (Radev et al., 2000). COGENT complements the default MEAD sentence scoring features with features based on findings from the empirical study. COGENT represents a hybrid approach integrating bottom-up (hypertext and content word density) and top-down (educational standards and gazetteer) features.

We model how human experts used external information sources with the educational standards feature. This feature leverages the text of the relevant AAAS Benchmarks and associated NSES. Each sentence receives a score based on its TFIDF similarity to the textual contents of these learning goals and educational standards.

We have developed a feature that reflects the large number of examples extracted by the experts. Earth science examples often refer to geographical locations and geological formations. The gazetteer feature checks named entities from each sentence against the Alexandria Digital Library (ADL) Gazetteer (Hill, 2000). A gazetteer is a geo-referencing resource containing location and type information about place-names. Each sentence receives a TFIDF score based on place-name term frequency and overall uniqueness in the gazetteer. Our assumption is that geographical locations with more unique names may be more pedagogically relevant.

Based on the intuition that the HTML structure of a resource reflects relevancy, we have developed the hypertext feature. This feature computes a sentence score directly proportional to the HTML heading level and inversely proportional to the relative paragraph number within a heading and to the relative sentence position within a paragraph.

To promote the extraction of sentences containing science concepts, we have developed the content word density feature. This feature computes the ratio of content to function words in a sentence. Function words are identified using a stopword list, and the feature only keeps sentences featuring more content words than function words.

We compute the final sentence score by adding the MEAD default feature scores (centroid and position) to the COGENT feature scores (educational standards, gazetteer, and hypertext). COGENT keeps sentences that pass the cut-off constraints, including the MEAD sentence length of 9 and COGENT content word density of 50%. The default MEAD cosine re-ranker eliminates redundant sentences. Since the experts used 5% of the total word count in the resources, we produce summaries of that same length.

5 Evaluation

We have evaluated COGENT by processing the 20 digital library resources used in the empirical study and comparing the output against the concepts identified by the experts. Three configurations are considered: Random, Default, and COGENT. The Random summary uses MEAD to extract random sentences. The Default summary uses the MEAD centroid, position and length default features. Finally, the COGENT summary extends MEAD with the COGENT features.

We use ROUGE (Lin, 2004) to assess summary quality using common n-gram counts and longest common subsequence (LCS) measures. We report on ROUGE-1 (unigrams), ROUGE-2 (bigrams), ROUGE W-1.2 (weighted LCS), and ROUGE-S* (skip bigrams) as they have been shown to correlate well with human judgments for longer multi-document summaries (Lin, 2004). Table 1 shows the results for recall (R), precision (P), and balanced f-measure (F).

		Random	Default	COGENT
R-1	**R**	0.4855	0.4976	0.6073
	P	0.5026	0.5688	0.6034
	F	0.4939	0.5308	0.6054
R-2	**R**	0.0972	0.1321	0.1907
	P	0.1006	0.1510	0.1895
	F	0.0989	0.1409	0.1901
R-W-1.2	**R**	0.0929	0.0951	0.1185
	P	0.1533	0.1733	0.1877
	F	0.1157	0.1228	0.1453
R-S*	**R**	0.2481	0.2620	0.3820
	P	0.2657	0.3424	0.3772
	F	0.2566	0.2969	0.3796

Table 1. Quality evaluation results

Table 1 indicates that COGENT consistently outperforms the Random and Default summaries. These results indicate the promise of our approach to generate extractive summaries of educational science content. Given our interest in generating a pedagogically effective domain knowledge map, we have also conducted a content-centric evaluation.

To characterize the COGENT summary contents, one of the authors manually constructed a summary corresponding to the best case output for an extractive summarizer. This Best Case summary comprises all the sentences from the resources that align to all the concepts selected by the experts. This summary comprises 621 sentences consisting of 13,116 words, or about a 9% word compression.

We use ROUGE-L to examine the union LCS between the reference and candidate summaries, thus capturing their linguistic surface structure similarity. We also use MEAD to report on cosine similarity. Table 2 shows the results for recall (R), precision (P), and balanced f-measure (F).

		Random (5%)	Default (5%)	COGENT (5%)	Best Case (9%)
R-L	**R**	0.4814	0.4919	0.6021	0.9669
	P	0.4982	0.5623	0.5982	0.6256
	F	0.4897	0.5248	0.6001	0.7597
Cosine		0.5382	0.6748	0.8325	0.9323

Table 2. Content evaluation results (word compression)

The ROUGE-L scores consistently indicate that the COGENT summary may be closer to the reference in linguistic surface structure than either the Random or Default summaries. Since the COGENT ROUGE-L recall score (R=0. 6021) is lower than the Best Case (R=0.9669), it is likely that COGENT may be extracting different sentences than those selected by the experts. Based on the high cosine similarity with the reference (0.8325), we hypothesize that COGENT may be selecting sentences that cover very similar concepts to those selected by the experts, but expressed differently.

Given the difference in word compression for the Best Case summary, we have performed an incremental analysis using the ROUGE-L measure shown in Figure 2.

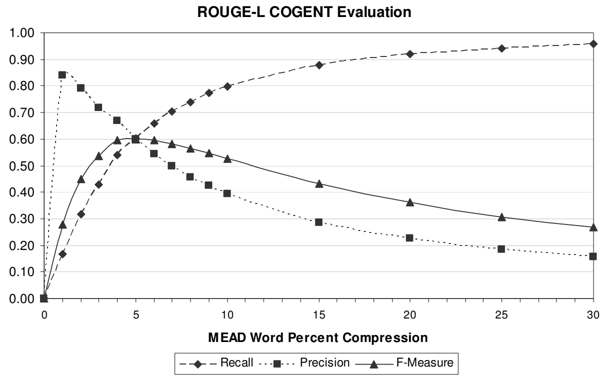

Figure 2. Incremental COGENT ROUGE-L analysis

Figure 2 indicates that COGENT can match the Best Case recall (R=0.9669) by generating a longer summary. For educational applications, lengthier summaries may be better suited for computational purposes, such as diagnosing student understanding, while shorter summaries may be more appropriate for display to the student.

6 Conclusions

COGENT extends MEAD based on strategies elicited from an empirical study with domain and instructional experts. Initial evaluation results indicate that COGENT holds promise for identifying important domain pedagogical concepts. We are exploring portability to other science education domains and machine learning techniques to connect concepts into a knowledge map. Automating the creation of inventories of pedagogically important concepts may represent an important step towards scalable intelligent tutoring systems.

Acknowledgements

This research is funded in part by the National Science Foundation under NSF IIS/ALT Award 0537194. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

T. Byrt, J. Bishop and J. B. Carlin. Bias, prevalence, and kappa. *Journal of Clinical Epidemiology*, *46*, *5* (1993), 423-429.

S. de la Chica, F. Ahmad, T. Sumner, J. H. Martin and K. Butcher. Computational foundations for personalizing instruction with digital libraries. *International Journal of Digital Libraries*, to appear in the Special Issue on Digital Libraries and Education.

S. Harabagiu and F. Lacatusu. Topic themes for multi-document summarization. In *Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Salvador, Brazil, 2005), 202-209.

H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, G. B. Wise and X. Zhang. Summarizing large document sets using concept-based clustering. In *Proc. of the Human Language Technology Conference 2002*, (San Diego, California, United States, 2002), 222-227.

L. L. Hill. Core elements of digital gazetteers: placenames, categories, and footprints. In *Proc. of the 4th European Conference on Digital Libraries*, (Lisbon, Portugal, 2000), 280-290.

C. D. Holley and D. F. Dansereau. *Spatial learning strategies: Techniques, applications, and related issues*. Academic Press, Orlando, Florida, 1984.

C. Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proc. of the Workshop on Text Summarization Branches Out*, (Barcelona, Spain, 2004).

D. Lin and P. Pantel. Concept discovery from text. In *Proc. of the 19th International Conference on Computational Linguistics*, (Taipei, Taiwan, 2002), 1-7.

National Research Council. *National Science Education Standards*. National Academy Press, Washington, DC, 1996.

R. Navigli and P. Velardi. Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, *30*, *2* (2004), 151-179.

A. M. O'Donnell, D. F. Dansereau and R. H. Hall. Knowledge maps as scaffolds for cognitive processing. *Educational Psychology Review*, *14*, *1* (2002), 71-86.

Project 2061. *Benchmarks for science literacy*. Oxford University Press, New York, New York, United States, 1993.

D. R. Radev, H. Jing and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proc. of the ANLP/NAACL 2000 Workshop on Summarization*, (2000), 21-30.

T. Sumner, S. Bhushan, F. Ahmad and Q. Gu. Designing a language for creating conceptual browsing interfaces for digital libraries. In *Proc. of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, (Houston, Texas, 2003), 258-260.

A. Zouaq, R. Nkambou and C. Frasson. Learning a domain ontology in the Knowledge Puzzle project. In *Proc. of the Fifth International Workshop on Ontologies and Semantic Web for E-Learning*, (Marina del Rey, California, 2007).