Towards a Framework for Abstractive Summarization of Multimodal Documents

Charles F. Greenbacker
Dept. of Computer & Information Sciences
University of Delaware
Newark, Delaware, USA
charlieg@cis.udel.edu

Abstract

We propose a framework for generating an abstractive summary from a semantic model of a multimodal document. We discuss the type of model required, the means by which it can be constructed, how the content of the model is rated and selected, and the method of realizing novel sentences for the summary. To this end, we introduce a metric called *information density* used for gauging the importance of content obtained from text and graphical sources.

1 Introduction

The automatic summarization of text is a prominent task in the field of natural language processing (NLP). While significant achievements have been made using statistical analysis and sentence extraction, "true abstractive summarization remains a researcher's dream" (Radev et al., 2002). Although existing systems produce high-quality summaries of relatively simple articles, there are limitations as to the types of documents these systems can handle.

One such limitation is the summarization of multimodal documents: no existing system is able to incorporate the non-text portions of a document (e.g., information graphics, images) into the overall summary. Carberry et al. (2006) showed that the content of information graphics is often not repeated in the article's text, meaning important information may be overlooked if the graphical content is not included in the summary. Systems that perform statistical analysis of text and extract sentences from the original article to assemble a summary cannot access the information contained in non-text components, let alone seamlessly combine that information with the extracted text. The problem is that information from the text and graphical components can only be integrated at the *conceptual* level, necessitating a semantic understanding of the underlying concepts.

Our proposed framework enables the generation of abstractive summaries from unified semantic models, regardless of the original format of the information sources. We contend that this framework is more akin to the human process of conceptual integration and regeneration in writing an abstract, as compared to the traditional NLP techniques of rating and extracting sentences to form a summary. Furthermore, this approach enables us to generate summary sentences about the information collected from graphical formats, for which there are no sentences available for extraction, and helps avoid the issues of coherence and ambiguity that tend to affect extraction-based summaries (Nenkova, 2006).

2 Related Work

Summarization is generally seen as a two-phase process: identifying the important elements of the document, and then using those elements to construct a summary. Most work in this area has focused on extractive summarization, assembling the summary from sentences representing the information in a document (Kupiec et al., 1995). Statistical methods are often employed to find key words and phrases (Witbrock and Mittal, 1999). Discourse structure (Marcu, 1997) also helps indicate the most important sentences. Various machine learning techniques have been applied (Aone et al., 1999; Lin, 1999), as well as approaches combining surface, content, rel-

Proceedings of the ACL-HLT 2011 Student Session, pages 75–80,
Portland, OR, USA 19-24 June 2011.

evance and event features (Wong et al., 2008).

However, a few efforts have been directed towards abstractive summaries, including the modification (i.e., editing and rewriting) of extracted sentences (Jing and McKeown, 1999) and the generation of novel sentences based on a deeper understanding of the concepts being described. Lexical chains, which capture relationships between related terms in a document, have shown promise as an intermediate representation for producing summaries (Barzilay and Elhadad, 1997). Our work shares similarities with the knowledge-based text condensation model of Reimer and Hahn (1988), as well as with Rau et al. (1989), who developed an information extraction approach for conceptual information summarization. While we also build a conceptual model, we believe our method of construction will produce a richer representation. Moreover, Reimer and Hahn did not actually produce a natural language summary, but rather a condensed text graph.

Efforts towards the summarization of multimodal documents have included naïve approaches relying on image captions and direct references to the image in the text (Bhatia et al., 2009), while content-based image analysis and NLP techniques are being combined for multimodal document indexing and retrieval in the medical domain (Névéol et al., 2009).

3 Method

Our method consists of the following steps: building the semantic model, rating the informational content, and generating a summary. We construct the semantic model in a knowledge representation based on typed, structured objects organized under a foundational ontology (McDonald, 2000). To analyze the text, we use Sparser,[1] a linguistically-sound, phrase structure-based chart parser with an extensive and extendible semantic grammar (McDonald, 1992). For the purposes of this proposal, we assume a relatively complete semantic grammar exists for the domain of documents to be summarized. In the prototype implementation (currently in progress), we are manually extending an existing grammar on an as-needed basis, with plans for large-scale learning of new rules and ontology definitions as future work. Projects like the Never-Ending Language Learner (Carlson et al., 2010) may enable us to induce these resources automatically.

Although our framework is general enough to cover any image type, as well as other modalities (e.g., audio, video), since image understanding research has not yet developed tools capable of extracting semantic content from every possible image, we must restrict our focus to a limited class of images for the prototype implementation. Information graphics, such as bar charts and line graphs, are commonly found in popular media (e.g., magazines, newspapers) accompanying article text. To integrate this graphical content, we use the SIGHT system (Demir et al., 2010b) which identifies the intended message of a bar chart or line graph along with other salient propositions conveyed by the graphic. Extending the prototype to incorporate other modalities would not entail a significant change to the framework. However, it would require adding a module capable of mapping the particular modality to its underlying message-level semantic content.

The next sections provide detail regarding the steps of our method, which will be illustrated on a short article from the May 29, 2006 edition of Businessweek magazine entitled, "Will Medtronic's Pulse Quicken?"[2] This particular article was chosen due to good coverage in the existing Sparser grammar for the business news domain, and because it appears in the corpus of multimodal documents made available by the SIGHT project.

3.1 Semantic Modeling

Figure 1 shows a high-level (low-detail) overview of the type of semantic model we can build using Sparser and SIGHT. This particular example models the article text (including title) and line graph from the Medtronic article. Each box represents an individual concept recognized in the document. Lines connecting boxes correspond to relationships between concepts. In the interest of space, the individual attributes of the model entries have been omitted from this diagram, but are available in Figure 2, which zooms into a fragment of the model showing the concepts that are eventually rated most salient (Section 3.2) and selected for inclusion in

[1] https://github.com/charlieg/Sparser

[2] Available at http://www.businessweek.com/magazine/content/06_22/b3986120.htm.

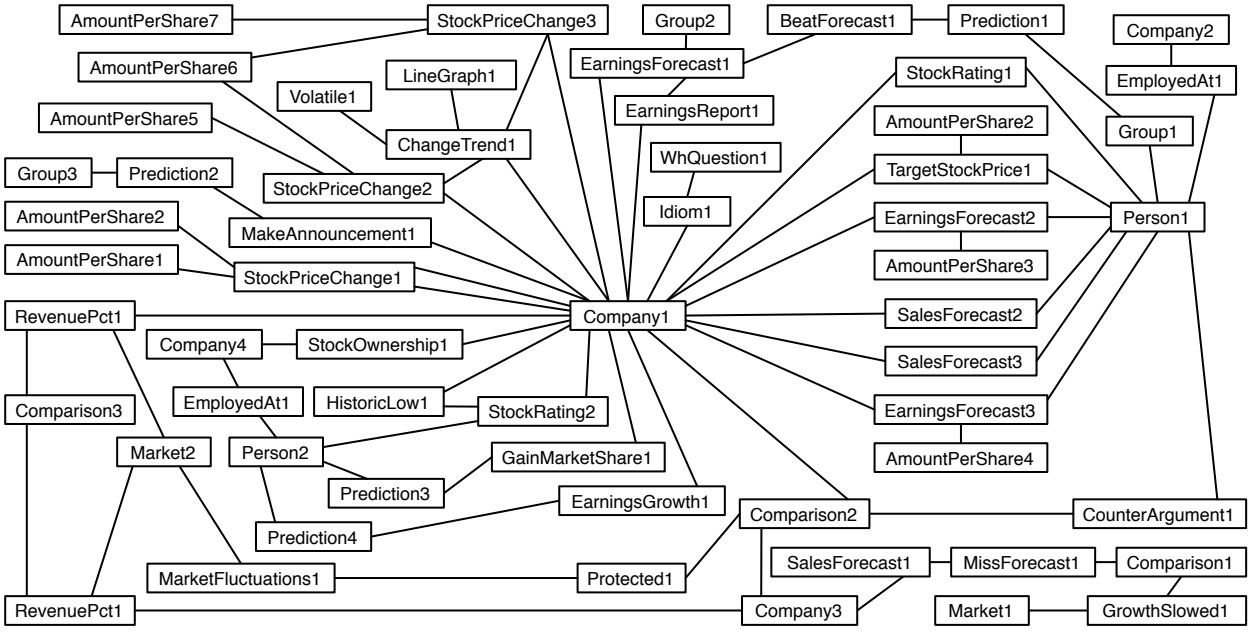

Figure 1: High-level overview of semantic model for Medtronic article.

the summary (Section 3.3). The top portion of each box in Figure 2 indicates the name of the conceptual category (with a number to distinguish between instances), the middle portion shows various attributes of the concept with their values, and the bottom portion contains some of the original phrasings from the text that were used to express these concepts (formally stored as a synchronous TAG) (McDonald and Greenbacker, 2010)). Attribute values in angle brackets (<>) are references to other concepts, hash symbols (#) refer to a concept or category that has not been instantiated in the current model, and each expression is preceded by a sentence tag (e.g., "P1S4" stands for "paragraph 1, sentence 4").

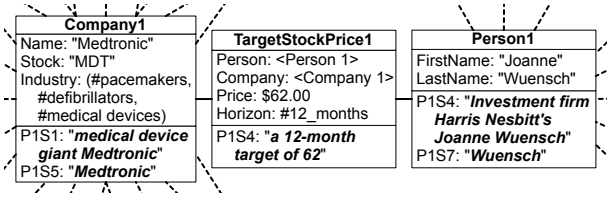

Figure 2: Detail of Figure 1 showing concepts rated most important and selected for inclusion in the summary.

As illustrated in this example, concepts conveyed by the graphics in the document can also be included in the semantic model. The overall intended message (ChangeTrend1) and additional propositions (Volatile1, StockPriceChange3, etc.) that SIGHT extracts from the line graph and deems important are added to the model produced by Sparser by simply inserting new concepts, filling slots for existing concepts, and creating new connections. This way, information gathered from both text and graphical sources can be integrated at the *conceptual* level regardless of the format of the source.

3.2 Rating Content

Once document analysis is complete and the semantic model has been built, we must determine which concepts conveyed by the document and captured in the model are most salient. Intuitively, the concepts containing the most information and having the most connections to other important concepts in the model are those we'd like to convey in the summary. We propose the use of an *information density* metric (ID) which rates a concept's importance based on a number of factors:[3]

- Completeness of attributes: the concept's filled-in slots (f) vs. its total slots (s) ["saturation level"], and the importance of the concepts (c_i) filling these slots [a recursive value]:

$$\frac{f}{s} * log(s) * \sum_{i=1}^{f} ID(c_i)$$

[3]The first three factors are similar to the dominant slot fillers, connectivity patterns, and frequency criteria described by Reimer and Hahn (1988).

- Number of connections/relationships (n) with other concepts (c_j), and the importance of these connected concepts [a recursive value]:

$$\sum_{j=1}^{n} ID(c_j)$$

- Number of expressions (e) realizing the concept in the current document

- Prominence based on document and rhetorical structure (W_D & W_R), and salience assessed by the graph understanding system (W_G)

Saturation refers to the level of completeness with which the knowledge base entry for a given concept is "filled-out" by information obtained from the document. As information is collected about a concept, the corresponding slots in its concept model entry are assigned values. The more slots that are filled, the more we know about a given instance of a concept. When all slots are filled, the model entry for that concept is "complete," at least as far as the ontological definition of the concept category is concerned. As saturation level is sensitive to the amount of detail in the ontology definition, this factor must be normalized by the number of attribute slots in its definition, thus $log(s)$ above.

In Figure 3 we can see an example of relative saturation level by comparing the attribute slots for Company2 with that of Company1 in Figure 2. Since the "Stock" slot is filled for Medtronic and remains empty for Harris Nesbitt, we say that the concept for Company1 is more saturated (i.e., more complete) than that of Company2.

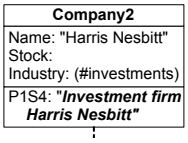

Figure 3: Detail of Figure 1 showing example concept with unfilled attribute slot.

Document and rhetorical structure (W_D and W_R) take into account the location of a concept within a document (e.g., mentioned in the title) and the use of devices highlighting particular concepts (e.g., juxtaposition) in computing the overall ID score. For the intended message and informational propositions conveyed by the graphics, the weights assigned by SIGHT are incorporated into ID as W_G.

After computing the ID of each concept, we will apply Demir's (2010a) graph-based ranking algorithm to select items for the summary. This algorithm is based on PageRank (Page et al., 1999), but with several changes. Beyond centrality assessment based on relationships between concepts, it also incorporates apriori importance nodes that enable us to capture concept completeness, number of expressions, and document and rhetorical structure. More importantly from a generation perspective, Demir's algorithm iteratively selects concepts one at a time, re-ranking the remaining items by increasing the weight of related concepts and discounting redundant ones. Thus, we favor concepts that ought to be conveyed together while avoiding redundancy.

3.3 Generating a Summary

After we determine which concepts are most important as scored by ID, the next step is to decide what to say about them and express these elements as sentences. Following the generation technique of McDonald and Greenbacker (2010), the expressions observed by the parser and stored in the model are used as the "raw material" for expressing the concepts and relationships. The two most important concepts as rated in the semantic model built from the Medtronic article would be Company1 ("Medtronic") and Person1 ("Joanne Wuensch," a stock analyst). To generate a single summary sentence for this document, we should try to find some way of expressing these concepts together using the available phrasings. Since there is no direct link between these two concepts in the model (see Figure 1), none of the collected phrasings can express both concepts at the same time. Instead, we need to find a third concept that provides a semantic link between Company1 and Person1. If multiple options are available, deciding which linking concept to use becomes a microplanning problem, with the choice depending on linguistic constraints and the relative importance of the applicable linking concepts.

In this example, a reasonable selection would be TargetStockPrice1 (see Figure 1). Combining original phrasings from all three concepts (via substitution and adjunction operations on the underlying TAG trees), along with a "built-in" realization inherited by the TargetStockPrice category (a subtype of Expectation – not shown in the figure), produces a

construction resulting in this final surface form:

> *Wuensch expects a 12-month target of 62 for medical device giant Medtronic.*

Thus, we generate novel sentences, albeit with some "recycled" expressions, to form an abstractive summary of the original document.

Studies have shown that nearly 80% of human-written summary sentences are produced by a cut-and-paste technique of reusing original sentences and editing them together in novel ways (Jing and McKeown, 1999). By reusing selected short phrases ("cutting") coupled together with generalized constructions ("pasting"), we can generate abstracts similar to human-written summaries.

The set of available expressions is augmented with numerous built-in schemas for realizing common relationships such as "is-a" and "has-a," as well as realizations inherited from other conceptual categories in the hierarchy. If the knowledge base persists between documents, storing the observed expressions and making them available for later use when realizing concepts in the same category, the variety of utterances we can generate is increased. With a sufficiently rich set of expressions, the reliance on straightforward "recycling" is reduced while the amount of paraphrasing and transformation is increased, resulting in greater novelty of production. By using ongoing parser observations to support the generation process, the more the system "reads," the better it "writes."

4 Evaluation

As an intermediate evaluation, we will rate the concepts stored in a model built only from text and use this rating to select sentences containing these concepts from the original document. These sentences will be compared to another set chosen by traditional extraction methods. Human judges will be asked to determine which set of sentences best captures the most important concepts in the document. This "checkpoint" will allow us to assess how well our system identifies the most salient concepts in a text.

The summaries ultimately generated as final output by our prototype system will be evaluated against summaries written by human authors, as well as summaries created by extraction-based systems and a baseline of selecting the first few sentences. For each comparison, participants will be asked to indicate a preference for one summary over another. We propose to use preference-strength judgment experiments testing multiple dimensions of preference (e.g., accuracy, clarity, completeness). Compared to traditional rating scales, this alternative paradigm has been shown to result in better evaluator self-consistency and high inter-evaluator agreement (Belz and Kow, 2010). This allows a larger proportion of observed variations to be accounted for by the characteristics of systems undergoing evaluation, and can result in a greater number of significant differences being discovered.

Automatic evaluation, though desirable, is likely unfeasible. As human-written summaries have only about 60% agreement (Radev et al., 2002), there is no "gold standard" to compare our output against.

5 Discussion

The work proposed herein aims to advance the state-of-the-art in automatic summarization by offering a means of generating abstractive summaries from a semantic model built from the original article. By incorporating concepts obtained from non-text components (e.g., information graphics) into the semantic model, we can produce unified summaries of multimodal documents, resulting in an abstract covering the entire document, rather than one that ignores potentially important graphical content.

Acknowledgments

This work was funded in part by the National Institute on Disability and Rehabilitation Research (grant #H133G080047). The author also wishes to thank Kathleen McCoy, Sandra Carberry, and David McDonald for their collaborative support.

References

Chinatsu Aone, Mary E. Okurowski, James Gorlinsky, and Bjornar Larsen. 1999. A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automated Text Summarization*. MIT Press.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *In Proceedings*

of the ACL Workshop on Intelligent Scalable Text Summarization, pages 10–17, Madrid, July. ACL.

Anja Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG 2010, pages 7–16, Trim, Ireland, July. ACL.

Sumit Bhatia, Shibamouli Lahiri, and Prasenjit Mitra. 2009. Generating synopses for document-element search. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 2003–2006, Hong Kong, November. ACM.

Sandra Carberry, Stephanie Elzer, and Seniz Demir. 2006. Information graphics: an untapped resource for digital libraries. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 581–588, Seattle, August. ACM.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the 24th Conference on Artificial Intelligence (AAAI 2010)*, pages 1306–1313, Atlanta, July. AAAI.

Seniz Demir, Sandra Carberry, and Kathleen F. McCoy. 2010a. A discourse-aware graph-based content-selection framework. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG 2010, pages 17–26, Trim, Ireland, July. ACL.

Seniz Demir, David Oliver, Edward Schwartz, Stephanie Elzer, Sandra Carberry, and Kathleen F. McCoy. 2010b. Interactive SIGHT into information graphics. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility*, W4A '10, pages 16:1–16:10, Raleigh, NC, April. ACM.

Hongyan Jing and Kathleen R. McKeown. 1999. The decomposition of human-written summary sentences. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 129–136, Berkeley, August. ACM.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 68–73, Seattle, July. ACM.

Chin-Yew Lin. 1999. Training a selection function for extraction. In *Proceedings of the 8th International Conference on Information and Knowledge Management*, CIKM '99, pages 55–62, Kansas City, November. ACM.

Daniel C. Marcu. 1997. *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*. Ph.D. thesis, University of Toronto, December.

David D. McDonald and Charles F. Greenbacker. 2010. 'If you've heard it, you can say it' - towards an account of expressibility. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG 2010, pages 185–190, Trim, Ireland, July. ACL.

David D. McDonald. 1992. An efficient chart-based algorithm for partial-parsing of unrestricted texts. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 193–200, Trento, March. ACL.

David D. McDonald. 2000. Issues in the representation of real texts: the design of KRISP. In Lucja M. Iwańska and Stuart C. Shapiro, editors, *Natural Language Processing and Knowledge Representation*, pages 77–110. MIT Press, Cambridge, MA.

Ani Nenkova. 2006. *Understanding the process of multi-document summarization: content selection, rewrite and evaluation*. Ph.D. thesis, Columbia University, January.

Aurélie Névéol, Thomas M. Deserno, Stéfan J. Darmoni, Mark Oliver Güld, and Alan R. Aronson. 2009. Natural language processing versus content-based image analysis for medical document retrieval. *Journal of the American Society for Information Science and Technology*, 60(1):123–134.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November. Previous number: SIDL-WP-1999-0120.

Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408.

Lisa F. Rau, Paul S. Jacobs, and Uri Zernik. 1989. Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*, 25(4):419 – 428.

Ulrich Reimer and Udo Hahn. 1988. Text condensation as knowledge base abstraction. In *Proceedings of the 4th Conference on Artificial Intelligence Applications*, CAIA '88, pages 338–344, San Diego, March. IEEE.

Michael J. Witbrock and Vibhu O. Mittal. 1999. Ultra-summarization: a statistical approach to generating highly condensed non-extractive summaries. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 315–316, Berkeley, August. ACM.

Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd Int'l Conference on Computational Linguistics*, COLING '08, pages 985–992, Manchester, August. ACL.