Toward Extractive Summarization of Multimodal Documents

Peng Wu and Sandra Carberry

Computer and Information Science Department
University of Delaware,
Newark, DE, 19716
{pwu,carberry}@cis.udel.edu

Abstract. Summarization research has focused on text, and relatively little attention has been given to the summarization of multimodal documents. If extractive summarization techniques are to be used on multimodal documents containing information graphics (bar charts, line graphs, etc.), then a strategy must be devised both for extracting the high-level content of the information graphics and for identifying where that content is relevant in the article's text. This paper gives an overview of our prior work on constructing a summary of an information graphic and presents our new research on methods for selecting paragraphs in a multimodal document that are most relevant to a constituent information graphic. The results demonstrate that our methods are far superior to possible baseline methods and that our work advances the use of extractive techniques for summarizing multimodal documents.

1 Introduction

Summarization research has focused on text, and little attention has been given to multimodal documents. For the most part, this has been due to the difficulty of identifying the content of non-textual components of a document and how this content relates to the document's text. We are addressing the summarization of multimodal documents that consist of text and information graphics, where an information graphic is defined as a non-pictorial graphic such as a bar chart or a line graph. As shown by [2], the message conveyed by an information graphic in popular media (such as newspapers and magazines, as opposed to scientific articles) is often not repeated in the article's text; furthermore, the graphic's caption often contains little or none of the graphic's primary intended message. Thus, information graphics in multimodal documents cannot be ignore.

In previous research[5], we developed a system for constructing a brief summary of information graphics that appear in popular media. One goal of our current research is to extend this work to the summarization of multimodal documents by inserting the graph's summary into the document's text and then applying traditional extractive summarization techniques to construct a summary of the entire document. Unfortunately, unlike scientific articles, the texts

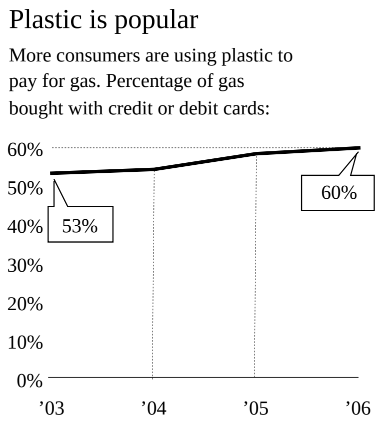

Fig. 1: A line graph from an article about consumer spending

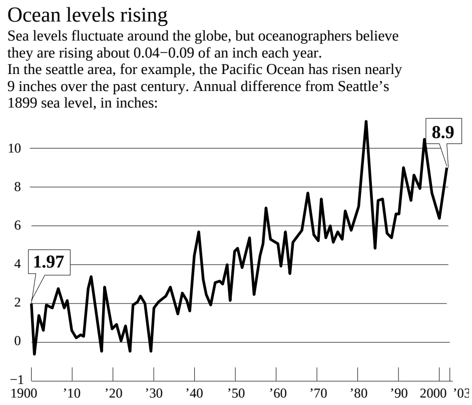

Fig. 2: A line graph from an article about global warming

of multimodal documents from popular media rarely refer explicitly to their information graphics and the graphics often do not appear adjacent to a relevant paragraph (or even on the same page). However, the graph’s summary must be inserted at a relevant point in the document. For example, the graph in Figure 1 is included in an article published in *USA Today* with the headline *“Paper or plastic? Answer might save at the pump”*. The most relevant paragraph within the article is the following:

– “More than three-quarters of the gas pumped in the USA is sold at convenience stores. In 2005, 58% of gas was bought using credit and debit cards. Retailers say that number has been climbing in 2006, Lenard says.”

But the paragraph closest to the line graph is the following:

– “But on a recent Monday morning, the restaurant owner from Edgemoor, S.C., took out his wallet, went into the gas station convenience store and paid with cash to take advantage of a 4-cent discount for cash customers.”

Thus extractive summarization of a multimodal document will lack coherence unless the appropriate placement of content from its information graphics can be identified.

This paper presents our implemented and evaluated methodology for identifying paragraphs in a document that are relevant to an information graphic’s content. Section 2 describes our prior work on summarizing information graphics and its relevance to extractive summarization of multimodal documents, along with other important applications of our research. Section 3 then presents our methodology for identifying paragraphs in a document’s text that are relevant to an information graphic, Section 4 discusses two examples processed by our system, and Section 5 discusses an evaluation of our methodology. Section 6 discusses related work, and Section 7 describes our future work on the project.

2 Extractive Summarization and Other Applications

Although abstractive summarization is the Holy Grail of summarization research, the state-of-the-art is extractive summarization in which important clauses or sentences are extracted from a document's text. The extracted text is then knitted together into a summary, with the pieces of text generally appearing in the same order as in the original article.

To produce a coherent summary of a multimodal document using extractive summarization techniques, two tasks must be addressed: 1) the construction of a summary of the content of the document's information graphics, and 2) the integration of the graphics' summaries into an overall summary of the document. In previous research, we devised an approach for constructing a summary of an information graphic appearing in popular media. For a line graph, a graph segmentation module first uses a support vector machine to segment the line graph into a sequence of visually distinguishable trends[12]. For example, the line graph in Figure 2 would be converted into two segments, a relatively flat segment from 1900 to 1930 and a rising segment from 1930 to 2003. Then the system extracts communicative signals from the graph, such as whether one bar is colored differently from the other bars, whether a point in a line graph is annotated with its value, or whether a bar label is mentioned in the caption. These communicative signals bring an entity into focus and are used as evidence in a Bayesian network that hypothesizes the graphic's intended message. For example, the intended message of the line graph in Figure 2 is that there is a changing trend in ocean levels — relatively stable between 1900 and 1930 and then rising from 1930 to 2003. The Bayesian network has been implemented for simple bar charts[7] and single line graphs[13]. Next content identification rules (developed from human subject experiments) are used to identify additional propositions that are salient in the graphic and relevant to the graphic's intended message, and these are combined to produce a brief summary of the graphic that is realized in natural language[5].

To produce a summary of a multimodal document containing information graphics, we propose to insert the graph's summary at a relevant point in the article's text and then use extractive summarization techniques to construct a summary of the entire document. But this requires that we identify where the graph's summary should be inserted in the article's text — i.e., which paragraph is most relevant to the information graphic.

In addition to facilitating the application of extractive summarization techniques to multimodal documents that contain information graphics, our work on identifying relevant paragraphs has several other important applications:

1. Our SIGHT system[6] provides blind individuals with access to multimodal documents. SIGHT works within Internet Explorer and uses JAWS screen-reading software. It reads the text of a document to the user; when it encounters an information graphic, it invokes our system to construct a summary of the graphic, which is then relayed to the user via speech. By identifying relevant paragraphs in the document, the effectiveness of the SIGHT system

could be improved by summarizing the graphics at the most appropriate points in the document.

2. We are investigating the indexing and retrieval of information graphics from a digital library. The retrieval methodology will involve a mixture model that takes into account the graphic's intended message, the graphic's textual component such as its caption, and the accompanying textual article. But articles are often long, and much of the article may not be relevant to the information graphic. Thus we hypothesize that our system will perform better if we can identify the paragraphs of the accompanying article that are most relevant to an information graphic and use only these paragraphs in the mixture model that ranks the graphic for retrieval in response to a user query.

3 Methodology for Identifying Relevant Paragraphs

To identify the paragraphs that are most relevant to an information graphic, Section 3.1 proposes a KL divergence based calculation which measures the similarity between the textual component of the line graph and the paragraphs. (The textual component of a line graph consists of three parts: the caption which is the main title for the information graphic, the description which is any additional text that elaborates on the caption, and the "text in graphic" which is any text appearing inside the graphic area.) Section 3.2 then proposes a second method that augments the textual component with words selected from a word list consisting of verbs and adjectives that commonly appear in documents containing information graphics and with the parameters of the intended message of a line graph. The first part of the augmented word list reflects domain-independent graphic content and thus captures words that might appear in a paragraph relevant to *any* information graphic; the parameters of the intended message reflect the line graph's specific content and thus might appear in a paragraph that is specific to this information graphic.

3.1 Method P-KL: KL divergence

Our basic algorithm uses Kullback-Leibler divergence to measure the similarity of two language models, one model for a paragraph in a document and one model for the information graphic's textual component. KL divergence has been widely used in natural language processing and text mining. It measures the difference between two distributions, either continuous or discrete and can be written as

$$D_{KL}(p||q) = \sum_{i \in V} p(i) log \frac{p(i)}{q(i)}$$

where i is the index of a word in vocabulary V, and p and q are two distributions of words. If p and q represent the same word distribution, $D_{KL}(p||q)$ will be 0. For our problem of identifying the relevant paragraphs, p is a smoothed word distribution built from the line graph's textual component, and q is another smoothed word distribution built from a paragraph in the corresponding document. Smoothing addresses the problem of instances with zero occurrences of a word in the word distribution, which will cause problems in computing the KL divergence. We assign the observed word its true word frequency and assign each unobserved word a low frequency (such as 0.01) and then normalize the word distribution. We rank the paragraphs by their KL divergence score from lowest to highest, since lower KL divergence scores indicate a higher similarity.

3.2 Method P-KLA: KL Divergence with Augmented Textual Component

Our first method only considered the textual component accompanying the line graph. But an information graphic consists of two parts: the textual part and the graphic part. Although the textual part can vary depending on the domain, much of the actual graphic is domain-independent and presents trends, rises or falls, results (higher or lower), or (in the case of bar charts) ranks or comparisons. Thus we decided to explore whether we could automatically extract a set of expansion words that are commonly used in paragraphs that are relevant to information graphics.

To construct this word set, we apply an iterative process in which we automatically identify pseudo relevant paragraphs for each information graphic, extract potential expansion words from the set of pseudo relevant paragraphs identified for all the information graphics, and then repeat the process after augmenting an information graphic's textual component with words from the expansion set. The process is repeated until the expansion word set does not change (convergence) or changes only minimally.

For each information graphic in our training set, we use KL divergence to identify three pseudo-relevant paragraphs in the document. This is similar to the pseudo relevance feedback technology used in information retrieval[15], except that the information retrieval process considers a single query whereas we are using a set of information graphics and associated documents to identify an expansion set that can be applied to all information graphics. If there are N information graphics, we produce a set of 3N relevant paragraphs. The next step is to extract a common word set from the set of pseudo-relevant paragraphs. We assume that the collection of pseudo relevant paragraphs was generated by two models, one producing words relevant to the information graphics and one producing words relevant to the topics of the documents. Let W_g represent the word frequency vector that generates words relevant to the information graphics, W_a represent the word frequency vector that generates words relevant to the domains of the articles, and W_p represent the word frequency vector of the pseudo-relevant paragraphs. We can compute W_p from the pseudo-relevant paragraphs, and we can estimate W_a as the word frequency vector for the entire articles. We want to compute W_g by filtering the components of W_a from W_p. This is similar to the work done by Widdows[11] on orthogonal negation of vector spaces. The problem can be formulated as follows:

1. $W_p = \alpha W_a + \beta W_g$ where $\alpha > 0$ and $\beta > 0$, which means the word frequency vector for the pseudo-relevant paragraphs is a linear combination of the background (topics) word frequency vector and the graphic word vector.
2. $< W_a, W_g >= 0$ which means the background word vector is orthogonal to the graph description word vector. We assume that when the author writes paragraphs that are unrelated to the graphic, he/she will not have the graphic words in mind. Therefore the graphic word vector is independent of the background word vector and these two share minimal information. Since we use a vector space model to represent W_a and W_g, orthogonality is obtained by assuming that these two word vectors have minimum similarity.
3. W_g is assumed to be a unit vector. Whether or not W_g is a unit vector is immaterial for our method, since we are interested only in the relative rank of the word frequencies, not their actual values. However, assuming that W_g is a unit vector gives us three equations in three unknowns (W_g, α, and β) which can be solved for W_g.

With these three assumptions, we obtain

$$\alpha = \frac{< W_p, W_a >}{< W_a, W_a >} \tag{1}$$

$$W_g = \text{normalized}\left(W_p - \frac{< W_p, W_a >}{< W_a, W_a >}\dot{W}_a\right) \tag{2}$$

After we compute W_g, we use WordNet to filter out words whose main sense is neither *verb* nor *adjective*, under the assumption that nouns will be relevant to the domains or topics of the graphs (and are thus *noise*) whereas we want a general set of words (such as *"increasing"*) that are typically used when writing about the data in graphs. To roughly estimate whether a word is predominantly a verb or adjective, we determine whether there are more verbg and adjective senses of the word in Wordnet than there are senses that are nouns.

We then rank the words in the filtered W_g by their frequency and select the k (we chose $k = 25$ in our experiments) most frequent words as our expansion word list. Since the textual components were used to identify pseudo-relevant paragraphs and then pseudo-relevant paragraphs (as opposed to truly relevant paragraphs) were used to construct the word list for expanding the textual components, the accuracy of both the pseudo-relevant paragraphs and the expansion word list are suspect. Thus we apply the two steps (identify pseudo-relevant paragraphs and then extract a word list for expanding the textual components) iteratively until convergence or minimal changes between iterations.

In addition, the parameters of an intended message capture domain-specific content of the graphic's communicative goal. For example, the intended message of the line graph in Figure 2 is ChangingTrend(1900, 1930, 2003) which means that the line graph conveys a changing trend in ocean levels over the period from 1900 to 2003 with the change from relatively stable to rising occurring in 1930. Thus we also added the parameters of the intended message to the augmented word list.

The result is the expansion word list used in method P-KLA. Because the textual component may be even shorter than the expansion word list, we won't add a word from the expansion word list to the textual component unless the compared paragraph also contains this word.

4 Examples

Consider first the graphic in Figure 1. It appeared in an article containing 38 paragraphs. As noted in Section 1, the closest paragraph has little relevance to the graphic. The most relevant paragraph is repeated below:

> *"More than three-quarters of the gas pumped in the USA is sold at convenience stores. In 2005, 58% of gas was bought using credit and debit cards. Retailers say that number has been climbing in 2006, Lenard says."*

Both of our human evaluators selected this paragraph as most relevant to the graphic, and our best performing method, P-KLA, did the same.

Now consider the graphic in Figure 2. This graphic appeared in an article on global warming containing 23 paragraphs. Not only does the paragraph closest to the graphic have little relevance to it, but also no paragraph in the article stands out as overwhelmingly most relevant to the graphic. In fact, the two evaluators selected three and four paragraphs respectively as most relevant, and not only did they differ on their top-ranked paragraph but they also had only one paragraph in common. Although the top-ranked paragraph identified by our best performing method, P-KLA, does not match the paragraph identified as best by either of the human evaluators, the top four paragraphs selected by P-KLA include the four distinct paragraphs identified as relevant by one of the human evaluators. This performance on such a difficult article indicates that our method can handle articles where the most relevant paragraph is not obvious.

5 Evaluation

5.1 The Dataset

We have compiled a dataset of 461 information graphics with full articles from multiple national sources such as *USA Today*, *Business Week*, *News Week*, *New York Times*, and *Wall Street Journal* and some local sources such as *The Wilmington News Journal*. At the time of submission of the final version of this paper, 66 graphs and articles had been analyzed by two human evaluators; thus they were held out as test data and the remainder were used as a training set to build the expansion word list discussed in Section 3.2. For the 66 articles in the test set, the two human evaluators identified paragraphs in each document that were relevant to its constituent information graphic and ranked them in terms of relevance. On average, Evaluator-1 selected 2 paragraphs and Evaluator-2 selected 1.71 paragraphs. For 63.6% of the graphs, the two evaluators agreed on the top ranked paragraph; this shows that in many cases, the most relevant paragraph is not obvious and that several possibilities exist.

5.2 Evaluation Criteria

Both of our methods(P-KL and P-KLA) processed the test set of 66 information graphics with accompanying articles, and each method produced a ranked list of the paragraphs in terms of relevance. We evaluated the results in several ways. For summarization, we want to insert the summary of the graphic at a coherent point in the article's text and then apply extractive summarization on the text. This leads to two evaluation criteria:

1. TOP: the method's success rate in selecting *the most relevant paragraph*, measured as how often the most relevant paragraph identified by the method matches one of the two evaluator's top-ranked paragraph.
2. COVERED: the method's success rate in selecting *a relevant paragraph*, measured as how often the most relevant paragraph identified by the method matches one of the paragraphs identified as relevant by the evaluators.

For our work on retrieving information graphics from a digital library, we want to use several paragraphs of the accompanying article in our mixture model[16] that will rank graphics for retrieval. Thus an appropriate evaluation criteria is normalized discounted cumulative gain (nDCG)[3]. The nDCG is between 0 and 1, and measures how well the rank-order of the paragraphs retrieved by our method agree with the rank-order of the paragraphs identified as relevant by our evaluators. nDCG is defined by the following formulas:

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{3}$$

$$\text{where } DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i)} \tag{4}$$

$$\text{and } IDCG_p \text{ is the highest possible } DCG_p \tag{5}$$

We set the cut off position at $p = 3$. The rel_i is the gain of retrieving a paragraph and the $\frac{1}{log_2(i)}$ is the discount according to its position i. The value of rel_i depends on p and the number of relevant paragraphs identified by the human evaluator. If the human evaluator identifies k paragraphs as relevant (where $k \leq p$), then rel_i=k if the i-th ranked paragraph by the system matches the top-ranked paragraph by the human evaluator and is equal to $k - 1$ or $k - 2$ if it matches the paragraph ranked second or third respectively by the human evaluator. Ranking a good paragraph higher gets less discount with the same gain, and ranking a better paragraph at the same position gets higher gain with the same discount.

5.3 Experimental Results

Figures 3 and 4 present the success rate for both of our methods for criteria TOP and COVERED, along with the success rates for two baseline methods: 1) selection of a random paragraph as most relevant, and 2) selection of the

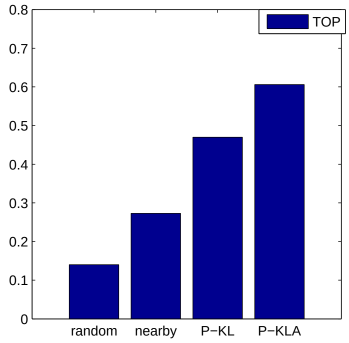

Fig. 3: Success rate in selecting the paragraph identified as most relevant by one of the two human evaluators

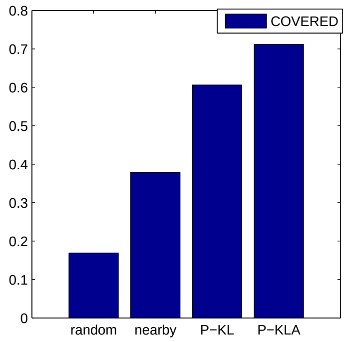

Fig. 4: Success rate in selecting a paragraph identified as relevant by one of the two human evaluators

paragraph that is closest to the information graphic. The results displayed in Figures 3 and 4 show that both of our methods outperform the baseline methods. P-KLA is a further improvement on P-KL. It selects the best paragraph in 60.6% of the test cases, and selects a relevant paragraph in 71.2% of the cases; for both criteria TOP and COVERED, P-KLA doubles or almost doubles the success rate of the baseline methods. The improvement of P-KLA over P-KL indicates that our expansion word list successfully expands the textual component with words pertinent to the graphic itself. A two-sided *student's t-test* shows that the improvements of P-KL over the baseline method and P-KLA over P-KL are both statistically significant at the 0.05 significance level.

Figure 5 presents the results of evaluating both methods in terms of the ranked order of their top three results using nDCG. We measured nDCG using each of the two evaluators as the ideal, and then averaged the results. (When comparing the two human evaluators against one another, their average nDCG is 0.69.) The baseline method in this evaluation is a random selection of three paragraphs from each document. The results in Figure 5 show that all of our methods outperformed the baseline. The best method is P-KLA which more than doubled the baseline method's nDCG. The improvement of P-KLA over P-KL is statistically significant at the 0.05 significance level.

5.4 Using sentence in addition to paragraph to improve the result

Though the paragraph based augmented KL-divergence method gave us satisfactory results, sometimes we consider a paragraph relevant only because there is a relevant sentence in the paragraph, without contribution from other sentences. We hypothesized that taking into consideration both the best sentence in a paragraph and the paragraph itself might further improve the result. We implemented another method named PM-KLA, which computes the final score for a paragraph as a weighted sum of the original score for the paragraph and

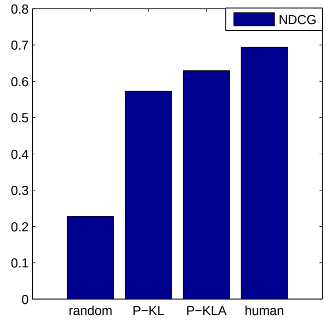

Fig. 5: nDCG scores for the algorithms and human evaluators

Criteria	P-KLA	PM-KLA
TOP	0.606	0.621
COVERED	0.712	0.727
nDCG	0.629	0.655

Table 1: Improved success rate of PM-KLA over P-KLA on three criteria

the score for the best sentence in the paragraph (the sentence with the lowest KL divergence from the augmented textual component).

$$Score_{final_p} = \lambda Score_{\text{best sentence}\in p} + (1-\lambda)Score_p$$

In our experiment, we arbitrarily chose $\lambda = 0.5$. Table 1 shows that the method (PM-KLA) has a higher success rate than P-KLA on both the TOP and COVERED criteria, and a higher nDCG score than P-KLA. However, these improvements are not statistically significant.

6 Related Work

Our work on identifying the paragraph that is most relevant to an information graphic in a multimodal document bears some similarity to the passage retrieval task in text retrieval[10] or question answering[4]. However, we are not doing passage retrieval based on a given query and there is only one document from which we must retrieve a relevant passage. This limits us from using multiple passages retrieved from multiple documents for the same query to improve the result with the relevance feedback technology[9].

Yu et al. [14] used a hierarchical clustering algorithm based on *tf-idf* to associate sentences from an abstract with images in biomedical articles. However, in scientific articles, the image is generally explicitly referred to by a sentence in the article. Thus their method used this referring sentence to identify words relevant to the image, which were likely to be repeated in the sentences of the abstract. In contrast, we are working with articles from popular media which generally have no such explicit reference to their information graphics; this makes our task more difficult.

A few research efforts have addressed multimodal summarization. Ahmad et al.[1] constructed a system for summarizing financial news and time series data. But instead of summarizing the time series data as text and inserting it into the article, they insert content from the articles into the time series data. Erol et al.[8] combines audio, video and a transcript of the recordings to produce a video summary of a meeting. They use *tf-idf* to identify significant words in the meeting transcript; then they use these words along with features such as intonation in the audio file and high motion in the video recording to identify significant events. These event segments are extracted from the video recording in the order of occurrence and spliced together to produce a video summary of the meeting. This differs from our work in that the different modalities are used only to extract segments from the video recording, whereas we must integrate information extracted from different modalities.

7 Conclusion and Future Work

Summarization is a difficult task, and a multimodal document compounds the problem. Our project's work[5] is the first to construct a summary of the knowledge conveyed by an information graphic, and we are extending this research to the summarization of multimodal documents. This paper addresses a key problem in extractive summarization of multimodal documents containing information graphics — namely, at what point in the document should the content of an information graphic be taken into account in the summary. We have presented methods for identifying the paragraph in the article's text that is most relevant to an information graphic, have analyzed the results produced by each method, and have shown that all of the methods perform far better than any baseline method that might be used. Not only can our best method be used to coherently integrate the content of an information graphic into a summary of a multimodal document, but it can also be used to select passages for use in a mixture model that ranks information graphics for retrieval in a digital library. In future work, we will explore how we might take the graphic's intended message into account when identifying relevant paragraphs and will investigate the quality of extractive summaries of multimodal documents using our approach.

8 Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1016916 and by the National Institute on Disability and Rehabilitation Research under Grant No. H133G080047. We also thank Matthew Stagitis from University of Delaware and Logan Peck from Millersville University for the corpus evaluation work.

References

1. Saif Ahmad, Paulo C F de Oliveira, and Khurshid Ahmad. Summarization of multimodal information. In *the proceeding of LREC 2004*, 2004.
2. Sandra Carberry, Stephanie Elzer, and Seniz Demir. Information graphics: An untapped resource for digital libraries. In *Proceedings of 9th International ACM SigIR Conference on Research and Development on Information Retrieval*, pages 581–588, 2006.
3. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, 2009.
4. Hang Cui, Renxu Sun, Keya Li, Min yen Kan, and Tat seng Chua. Question answering passage retrieval using dependency relations. In *Proceedings of 8th International ACM SigIR Conference on Research and Development on Information Retrieval*, pages 400–407. ACM Press, 2005.
5. Seniz Demir, Sandra Carberry, and Kathleen McCoy. Generating textual summaries of information graphics. In *Proceedings of the International Conference on Natural Language Generation*, pages 7–15, 2008.
6. Seniz Demir, David Oliver, Edward Schwartz, Stephanie Elzer, Sandra Carberry, Kathleen McCoy, and Daniel Chester. Interactive sight: Textual access to simple bar charts. *New Review of Hypermedia and Multimedia*, 16(3):245–279, 2010.
7. Stephanie Elzer, Sandra Carberry, and Ingrid Zukerman. The automated understanding of simple bar charts. *Artificial Intelligence*, 175:526–555, 2011.
8. B. Erol, D.-S. Lee, and J. Hull. Multimodal summarization of meeting recordings. In *Proceedings of International Conference on Multimedia and Expo*, vol. 3, pages 25–28, 2003.
9. Xiaoyan Li and Zhigang Zhu. Enhancing relevance models with adaptive passage retrieval. In *Advances in Information Retrieval*, pages 463–471, 2008.
10. Xiaoyong Liu and W. Bruce Croft. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management CIKM '02*, 2002.
11. Dominic Widdows. Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 136–143, 2003.
12. Peng Wu, Sandra Carberry, and Stephanie Elzer. Segmenting line graphs into trends. In *Proceedings of the Twelth International Conference on Artificial Intelligence*, pages 697–703, 2010.
13. Peng Wu, Sandra Carberry, Stephanie Elzer, and Daniel Chester. Recognizing the intended message of line graphs. In *Proceedings of the International Conference on the Theory and Applications of Diagrams*, pages 220–234, 2010.
14. Hong Yu and Minsuk Lee. Accessing bioscience images from abstract sentences. *Bioinformatics*, 22(14):547–556, 2006.
15. Chengxiang Zhai. *Statistical Language Models for Information Retrieval*. Morgan and Claypool Publishers, December 2008.
16. Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 403–410, 2001.