

Paragraph-, word-, and coherence-based approaches to sentence ranking: A comparison of algorithm and human performance

Florian WOLF

Massachusetts Institute of Technology
MIT NE20-448, 3 Cambridge Center
Cambridge, MA 02139, USA
fwolf@mit.edu

Edward GIBSON

Massachusetts Institute of Technology
MIT NE20-459, 3 Cambridge Center
Cambridge, MA 02139, USA
egibson@mit.edu

Abstract

Sentence ranking is a crucial part of generating text summaries. We compared human sentence rankings obtained in a psycholinguistic experiment to three different approaches to sentence ranking: A simple paragraph-based approach intended as a baseline, two word-based approaches, and two coherence-based approaches. In the paragraph-based approach, sentences in the beginning of paragraphs received higher importance ratings than other sentences. The word-based approaches determined sentence rankings based on relative word frequencies (Luhn (1958); Salton & Buckley (1988)). Coherence-based approaches determined sentence rankings based on some property of the coherence structure of a text (Marcu (2000); Page et al. (1998)). Our results suggest poor performance for the simple paragraph-based approach, whereas word-based approaches perform remarkably well. The best performance was achieved by a coherence-based approach where coherence structures are represented in a non-tree structure. Most approaches also outperformed the commercially available MSWord summarizer.

1 Introduction

Automatic generation of text summaries is a natural language engineering application that has received considerable interest, particularly due to the ever-increasing volume of text information available through the internet. The task of a human generating a summary generally involves three subtasks (Brandow et al. (1995); Mitra et al. (1997)): (1) understanding a text; (2) ranking text pieces (sentences, paragraphs, phrases, etc.) for importance; (3) generating a new text (the summary). Like most approaches to summarization, we are concerned with the second subtask (e.g. Carlson et al. (2001); Goldstein et al. (1999); Gong & Liu (2001); Jing et al. (1998);

Luhn (1958); Mitra et al. (1997); Sparck-Jones & Sakai (2001); Zechner (1996)). Furthermore, we are concerned with obtaining generic rather than query-relevant importance rankings (cf. Goldstein et al. (1999), Radev et al. (2002) for that distinction).

We evaluated different approaches to sentence ranking against human sentence rankings. To obtain human sentence rankings, we asked people to read 15 texts from the Wall Street Journal on a wide variety of topics (e.g. economics, foreign and domestic affairs, political commentaries). For each of the sentences in the text, they provided a ranking of how important that sentence is with respect to the content of the text, on an integer scale from 1 (not important) to 7 (very important).

The approaches we evaluated are a simple paragraph-based approach that serves as a baseline, two word-based algorithms, and two coherence-based approaches¹. We furthermore evaluated the MSWord summarizer.

2 Approaches to sentence ranking

2.1 Paragraph-based approach

Sentences at the beginning of a paragraph are usually more important than sentences that are further down in a paragraph, due in part to the way people are instructed to write. Therefore, probably the simplest approach conceivable to sentence ranking is to choose the first sentences of each

¹ We did not use any machine learning techniques to boost performance of the algorithms we tested. Therefore performance of the algorithms tested here will almost certainly be below the level of performance that could be reached if we had augmented the algorithms with such techniques (e.g. Carlson et al. (2001)). However, we think that a comparison between ‘bare-bones’ algorithms is viable because it allows to see how performance differs due to different basic approaches to sentence ranking, and not due to potentially different effects of different machine learning algorithms on different basic approaches to sentence ranking. In future research we plan to address the impact of machine learning on the algorithms tested here.

paragraph as important, and the other sentences as not important. We included this approach merely as a simple baseline.

2.2 Word-based approaches

Word-based approaches to summarization are based on the idea that discourse segments are important if they contain “important” words. Different approaches have different definitions of what an important word is. For example, Luhn (1958), in a classic approach to summarization, argues that sentences are more important if they contain many significant words. Significant words are words that are not in some predefined stoplist of words with high overall corpus frequency². Once significant words are marked in a text, clusters of significant words are formed. A cluster has to start and end with a significant word, and fewer than n insignificant words must separate any two significant words (we chose $n = 3$, cf. Luhn (1958)). Then, the weight of each cluster is calculated by dividing the square of the number of significant words in the cluster by the total number of words in the cluster. Sentences can contain multiple clusters. In order to compute the weight of a sentence, the weights of all clusters in that sentence are added. The higher the weight of a sentence, the higher is its ranking.

A more recent and frequently used word-based method used for text piece ranking is *tf.idf* (e.g. Manning & Schuetze (2000); Salton & Buckley (1988); Sparck-Jones & Sakai (2001); Zechner (1996)). The *tf.idf* measure relates the frequency of words in a text piece, in the text, and in a collection of texts respectively. The intuition behind *tf.idf* is to give more weight to sentences that contain terms with high frequency in a document but low frequency in a reference corpus. Figure 1 shows a formula for calculating *tf.idf*, where ds_{ij} is the *tf.idf* weight of sentence i in document j , n_{si} is the number of words in sentence i , k is the k th word in sentence i , tf_{jk} is the frequency of word k in document j , n_d is the number of documents in the reference corpus, and df_k is the number of documents in the reference corpus in which word k appears.

$$ds_{ij} = \sum_{k=1}^{n_{si}} tf_{jk} \cdot \log \left(\frac{n_d}{df_k} \right)$$

Figure 1. Formula for calculating *tf.idf* (Salton & Buckley (1988)).

² Instead of stoplists, *tf.idf* values have also been used to determine significant words (e.g. Buyukkokten et al. (2001)).

We compared both Luhn (1958)’s measure and *tf.idf* scores to human rankings of sentence importance. We will show that both methods performed remarkably well, although one coherence-based method performed better.

2.3 Coherence-based approaches

The sentence ranking methods introduced in the two previous sections are solely based on layout or on properties of word distributions in sentences, texts, and document collections. Other approaches to sentence ranking are based on the informational structure of texts. With informational structure, we mean the set of informational relations that hold between sentences in a text. This set can be represented in a graph, where the nodes represent sentences, and labeled directed arcs represent informational relations that hold between the sentences (cf. Hobbs (1985)). Often, informational structures of texts have been represented as trees (e.g. Carlson et al. (2001), Corston-Oliver (1998), Mann & Thompson (1988), Ono et al. (1994)). We will present one coherence-based approach that assumes trees as a data structure for representing discourse structure, and one approach that assumes less constrained graphs. As we will show, the approach based on less constrained graphs performs better than the tree-based approach when compared to human sentence rankings.

3 Coherence-based summarization revisited

This section will discuss in more detail the data structures we used to represent discourse structure, as well as the algorithms used to calculate sentence importance, based on discourse structures.

3.1 Representing coherence structures

3.1.1 Discourse segments

Discourse segments can be defined as non-overlapping spans of prosodic units (Hirschberg & Nakatani (1996)), intentional units (Grosz & Sidner (1986)), phrasal units (Lascarides & Asher (1993)), or sentences (Hobbs (1985)). We adopted a sentence unit-based definition of discourse segments for the coherence-based approach that assumes non-tree graphs. For the coherence-based approach that assumes trees, we used Marcu (2000)’s more fine-grained definition of discourse segments because we used the discourse trees from Carlson et al. (2002)’s database of coherence-annotated texts.

3.1.2 Kinds of coherence relations

We assume a set of coherence relations that is similar to that of Hobbs (1985). Below are examples of each coherence relation.

- (1) Cause-Effect
[There was bad weather at the airport]_a [and so our flight got delayed.]_b
- (2) Violated Expectation
[The weather was nice]_a [but our flight got delayed.]_b
- (3) Condition
[If the new software works,]_a [everyone will be happy.]_b
- (4) Similarity
[There is a train on Platform A.]_a [There is another train on Platform B.]_b
- (5) Contrast
[John supported Bush]_a [but Susan opposed him.]_b
- (6) Elaboration
[A probe to Mars was launched this week.]_a [The European-built ‘Mars Express’ is scheduled to reach Mars by late December.]_b
- (7) Attribution
[John said that]_a [the weather would be nice tomorrow.]_b
- (8) Temporal Sequence
[Before he went to bed,]_a [John took a shower.]_b

Cause-effect, violated expectation, condition, elaboration, temporal sequence, and attribution are asymmetrical or directed relations, whereas *similarity, contrast, and temporal sequence* are symmetrical or undirected relations (Mann & Thompson, 1988; Marcu, 2000). In the non-tree-based approach, the directions of asymmetrical or directed relations are as follows: cause → effect for *cause-effect*; cause → absent effect for *violated expectation*; condition → consequence for *condition*; elaborating → elaborated for *elaboration*, and source → attributed for *attribution*. In the tree-based approach, the asymmetrical or directed relations are between a more important discourse segment, or a Nucleus, and a less important discourse segment, or a Satellite (Marcu (2000)). The Nucleus is the equivalent of the arc destination, and the Satellite is the equivalent of the arc origin in the non-tree-based approach. The symmetrical or undirected relations are between two discourse elements of equal importance, or two Nuclei. Below we will explain how the difference between Satellites and Nuclei is considered in tree-based sentence rankings.

3.1.3 Data structures for representing discourse coherence

As mentioned above, we used two alternative representations for discourse structure, tree- and

non-tree based. In order to illustrate both data structures, consider (9) as an example:

- (9) Example text
0. Susan wanted to buy some tomatoes.
 1. She also tried to find some basil.
 2. The basil would probably be quite expensive at this time of the year.

Figure 2 shows one possible tree representation of the coherence structure of (9)³. *Sim* represents a *similarity* relation, and *elab* an *elaboration* relation. Furthermore, nodes with a “Nuc” subscript are Nuclei, and nodes with a “Sat” subscript are Satellites.

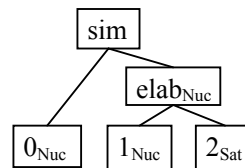


Figure 2. Coherence tree for (9).

Figure 3 shows a non-tree representation of the coherence structure of (9). Here, the heads of the arrows represent the directionality of a relation.

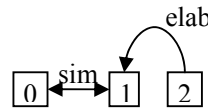


Figure 3. Non-tree coherence graph for (9).

3.2 Coherence-based sentence ranking

This section explains the algorithms for the tree- and the non-tree-based sentence ranking approach.

3.2.1 Tree-based approach

We used Marcu (2000)’s algorithm to determine sentence rankings based on tree discourse structures. In this algorithm, sentence salience is determined based on the tree level of a discourse segment in the coherence tree. Figure 4 shows Marcu (2000)’s algorithm, where $r(s,D,d)$ is the rank of a sentence s in a discourse tree D with depth d . Every node in a discourse tree D has a promotion set $promotion(D)$, which is the union of all Nucleus children of that node. Associated with every node in a discourse tree D is also a set of parenthetical nodes $parentheticals(D)$ (for example, in “Mars – half the size of Earth – is red”, “half the size of earth” would be a parenthetical node in a discourse tree). Both $promotion(D)$ and $parentheticals(D)$ can be empty sets. Furthermore, each node has a left subtree,

³ Another possible tree structure might be (elab (par (0 1) 2)).

$lc(D)$, and a right subtree, $rc(D)$. Both $lc(D)$ and $rc(D)$ can also be empty.

$$r(s, D, d) = \begin{cases} 0 & \text{if } D \text{ is NIL,} \\ d & \text{if } s \in \text{promotion}(D), \\ d-1 & \text{if } s \in \text{parentheticals}(D), \\ \max(r(s, lc(D), d-1), \\ r(s, rc(D), d-1)) & \text{otherwise} \end{cases}$$

Figure 4. Formula for calculating coherence-tree-based sentence rank (Marcu (2000)).

The discourse segments in Carlson et al. (2002)’s database are often sub-sentential. Therefore, we had to calculate sentence rankings from the rankings of the discourse segments that form the sentence under consideration. We did this by calculating the average ranking, the minimal ranking, and the maximal ranking of all discourse segments in a sentence. Our results showed that choosing the minimal ranking performed best, followed by the average ranking, followed by the maximal ranking (cf. Section 4.4).

3.2.2 Non-tree-based approach

We used two different methods to determine sentence rankings for the non-tree coherence graphs⁴. Both methods implement the intuition that sentences are more important if other sentences relate to them (Sparck-Jones (1993)).

The first method consists of simply determining the in-degree of each node in the graph. A node represents a sentence, and the in-degree of a node represents the number of sentences that relate to that sentence.

The second method uses Page et al. (1998)’s PageRank algorithm, which is used, for example, in the Google™ search engine. Unlike just determining the in-degree of a node, PageRank takes into account the importance of sentences that relate to a sentence. PageRank thus is a recursive algorithm that implements the idea that the more important sentences relate to a sentence, the more important that sentence becomes. Figure 5 shows how PageRank is calculated. PR_n is the PageRank of the current sentence, PR_{n-1} is the PageRank of the sentence that relates to sentence n , o_{n-1} is the out-degree of sentence $n-1$, and α is a damping parameter that is set to a value between 0 and 1. We report results for α set to 0.85 because this is a value often used in applications of PageRank (e.g. Ding et al. (2002); Page et al. (1998)). We also

calculated PageRanks for α set to values between 0.05 and 0.95, in increments of 0.05; changing α did not affect performance.

$$PR_n = 1 - \alpha + \alpha \frac{PR_{n-1}}{o_{n-1}}$$

Figure 5. Formula for calculating PageRank (Page et al. (1998)).

4 Experiments

In order to test algorithm performance, we compared algorithm sentence rankings to human sentence rankings. This section describes the experiments we conducted. In Experiment 1, the texts were presented with paragraph breaks; in Experiment 2, the texts were presented without paragraph breaks. This was done to control for the effect of paragraph information on human sentence rankings.

4.1 Materials for the coherence-based approaches

In order to test the tree-based approach, we took coherence trees for 15 texts from a database of 385 texts from the Wall Street Journal that were annotated for coherence (Carlson et al. (2002)). The database was independently annotated by six annotators. Inter-annotator agreement was determined for six pairs of two annotators each, resulting in kappa values (Carletta (1996)) ranging from 0.62 to 0.82 for the whole database (Carlson et al. (2003)). No kappa values for just the 15 texts we used were available.

For the non-tree based approach, we used coherence graphs from a database of 135 texts from the Wall Street Journal and the AP Newswire, annotated for coherence. Each text was independently annotated by two annotators. For the 15 texts we used, kappa was 0.78, for the whole database, kappa was 0.84.

4.2 Experiment 1: With paragraph information

15 participants from the MIT community were paid for their participation. All were native speakers of English and were naïve as to the purpose of the study (i.e. none of the subjects was familiar with theories of coherence in natural language, for example).

Participants were asked to read 15 texts from the Wall Street Journal, and, for each sentence in each text, to provide a ranking of how important that sentence is with respect to the content of the text, on an integer scale from 1 to 7 (1 = not important; 7 = very important). The texts were selected so

⁴ Neither of these methods could be implemented for coherence trees since Marcu (2000)’s tree-based algorithm assumes binary branching trees. Thus, the in-degree for all non-terminal nodes is always 2.

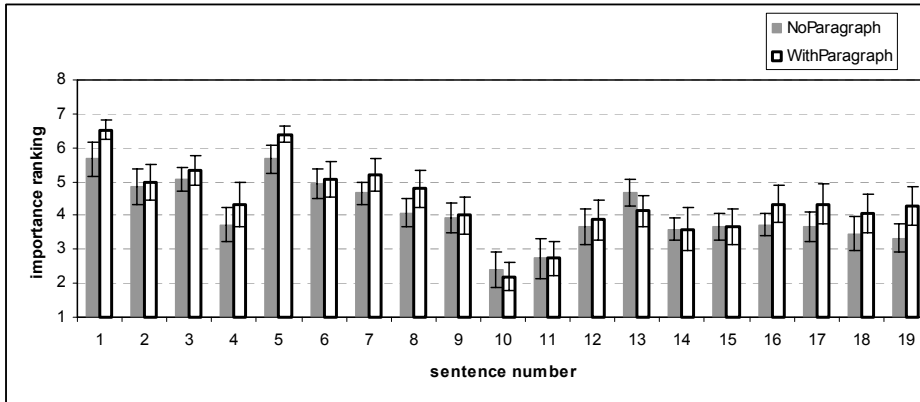


Figure 6. Human ranking results for one text (wsj_1306).

that there was a coherence tree annotation available in Carlson et al. (2002)’s database. Text lengths for the 15 texts we selected ranged from 130 to 901 words (5 to 47 sentences); average text length was 442 words (20 sentences), median was 368 words (16 sentences). Additionally, texts were selected so that they were about as diverse topics as possible.

The experiment was conducted in front of personal computers. Texts were presented in a web browser as one webpage per text; for some texts, participants had to scroll to see the whole text. Each sentence was presented on a new line. Paragraph breaks were indicated by empty lines; this was pointed out to the participants during the instructions for the experiment.

4.3 Experiment 2: Without paragraph information

The method was the same as in Experiment 1, except that texts in Experiment 2 did not include paragraph information. Each sentence was presented on a new line. None of the 15 participants who participated in Experiment 2 had participated in Experiment 1.

4.4 Results of the experiments

Human sentence rankings did not differ significantly between Experiment 1 and Experiment 2 for any of the 15 texts (all $F_s < 1$). This suggests that paragraph information does not have a big effect on human sentence rankings, at least not for the 15 texts that we examined. Figure 6 shows the results from both experiments for one text.

We compared human sentence rankings to different algorithmic approaches. The paragraph-based rankings do not provide scaled importance rankings but only “important” vs. “not important”. Therefore, in order to compare human rankings to the paragraph-based baseline approach, we

calculated point biserial correlations (cf. Bortz (1999)). We obtained significant correlations between paragraph-based rankings and human rankings only for one of the 15 texts.

All other algorithms provided scaled importance rankings. Many evaluations of scalable sentence ranking algorithms are based on precision/recall/F-scores (e.g. Carlson et al. (2001); Ono et al. (1994)). However, Jing et al. (1998) argue that such measures are inadequate because they only distinguish between hits and misses or false alarms, but do not account for a degree of agreement. For example, imagine a situation where the human ranking for a given sentence is “7” (“very important”) on an integer scale ranging from 1 to 7, and Algorithm A gives the same sentence a ranking of “7” on the same scale, Algorithm B gives a ranking of “6”, and Algorithm C gives a ranking of “2”. Intuitively, Algorithm B, although it does not reach perfect performance, still performs better than Algorithm C. Precision/recall/F-scores do not account for that difference and would rate Algorithm A as “hit” but Algorithm B as well as Algorithm C as “miss”. In order to collect performance measures that are more adequate to the evaluation of scaled importance rankings, we computed Spearman’s rank correlation coefficients. The rank correlation coefficients were corrected for tied ranks because in our rankings it was possible for more than one sentence to have the same importance rank, i.e. to have tied ranks (Horn (1942); Bortz (1999)).

In addition to evaluating word-based and coherence-based algorithms, we evaluated one commercially available summarizer, the MSWord summarizer, against human sentence rankings. Our reason for including an evaluation of the MSWord summarizer was to have a more useful baseline for scalable sentence rankings than the paragraph-based approach provides.

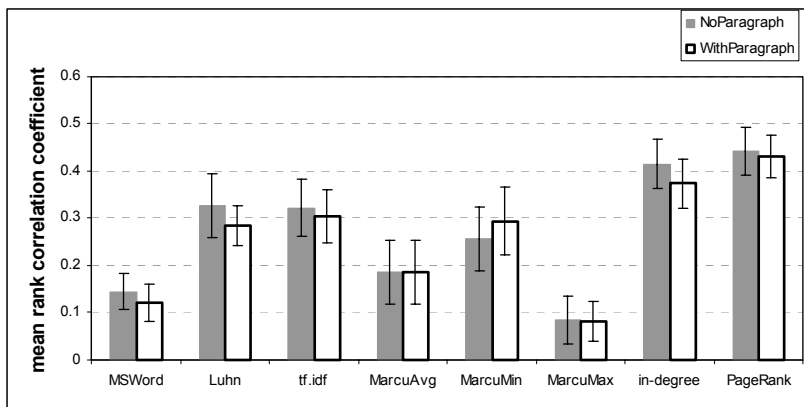


Figure 7. Average rank correlations of algorithm and human sentence rankings.

Figure 7 shows average rank correlations (ρ_{avg}) of each algorithm and human sentence ranking for the 15 texts. *MarcuAvg* refers to the version of Marcu (2000)’s algorithm where we calculated sentence rankings as the average of the rankings of all discourse segments that constitute that sentence; for *MarcuMin*, sentence rankings were the minimum of the rankings of all discourse segments in that sentence; for *MarcuMax* we selected the maximum of the rankings of all discourse segments in that sentence.

Figure 7 shows that the MSWord summarizer performed numerically worse than most other algorithms, except *MarcuMin*. Figure 7 also shows that PageRank performed numerically better than all other algorithms. Performance was significantly better than most other algorithms (*MSWord, NoParagraph*: $F(1,28) = 21.405$, $p = 0.0001$; *MSWord, WithParagraph*: $F(1,28) = 26.071$, $p = 0.0001$; *Luhn, WithParagraph*: $F(1,28) = 5.495$, $p = 0.026$; *MarcuAvg, NoParagraph*: $F(1,28) = 9.186$, $p = 0.005$; *MarcuAvg, WithParagraph*: $F(1,28) = 9.097$, $p = 0.005$; *MarcuMin, NoParagraph*: $F(1,28) = 4.753$, $p = 0.038$; *MarcuMax, NoParagraph*: $F(1,28) = 24.633$, $p = 0.0001$; *MarcuMax, WithParagraph*: $F(1,28) = 31.430$, $p = 0.0001$). Exceptions are *Luhn, NoParagraph* ($F(1,28) = 1.859$, $p = 0.184$); *tf.idf, NoParagraph* ($F(1,28) = 2.307$, $p = 0.14$); *MarcuMin, WithParagraph* ($F(1,28) = 2.555$, $p = 0.121$). The difference between *PageRank* and *tf.idf, WithParagraph* was marginally significant ($F(1,28) = 3.113$, $p = 0.089$).

As mentioned above, human sentence rankings did not differ significantly between Experiment 1 and Experiment 2 for any of the 15 texts (all $F_s < 1$). Therefore, in order to lend more power to our statistical tests, we collapsed the data for each text for the *WithParagraph* and the *NoParagraph* condition, and treated them as one experiment. Figure 8 shows that when the data from Experiments 1 and 2 are collapsed, *PageRank*

performed significantly better than all other algorithms except *in-degree* (two-tailed t-test results: *MSWord*: $F(1, 58) = 48.717$, $p = 0.0001$; *Luhn*: $F(1,58) = 6.368$, $p = 0.014$; *tf.idf*: $F(1,58) = 5.522$, $p = 0.022$; *MarcuAvg*: $F(1,58) = 18.922$, $p = 0.0001$; *MarcuMin*: $F(1,58) = 7.362$, $p = 0.009$; *MarcuMax*: $F(1,58) = 56.989$, $p = 0.0001$; *in-degree*: $F(1,58) < 1$).

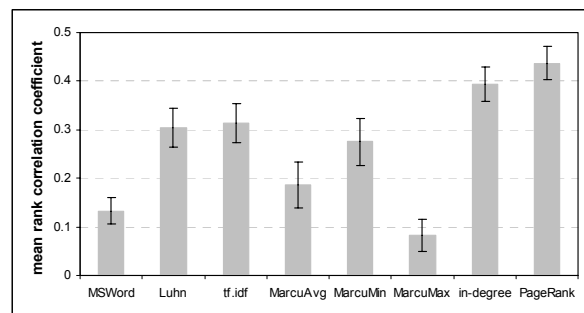


Figure 8. Average rank correlations of algorithm and human sentence rankings with collapsed data.

5 Conclusion

The goal of this paper was to evaluate the results of three different kinds of sentence ranking algorithms and one commercially available summarizer. In order to evaluate the algorithms, we compared their sentence rankings to human sentence rankings of fifteen texts of varying length from the Wall Street Journal.

Our results indicated that a simple paragraph-based algorithm that was intended as a baseline performed very poorly, and that word-based and some coherence-based algorithms showed the best performance. The only commercially available summarizer that we tested, the MSWord summarizer, showed worse performance than most other algorithms. Furthermore, we found that a coherence-based algorithm that uses *PageRank* and takes non-tree coherence graphs as input performed better than most versions of a

coherence-based algorithm that operates on coherence trees. When data from Experiments 1 and 2 were collapsed, the *PageRank* algorithm performed significantly better than all other algorithms, except the coherence-based algorithm that uses in-degrees of nodes in non-tree coherence graphs.

References

- Jürgen Bortz. 1999. *Statistik für Sozialwissenschaftler*. Berlin: Springer Verlag.
- Ronald Brandow, Karl Mitze, & Lisa F Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5), 675-685.
- Orkut Buyukkokten, Hector Garcia-Molina, & Andreas Paepcke. 2001. *Seeing the whole in parts: Text summarization for web browsing on handheld devices*. Paper presented at the 10th International WWW Conference, Hong Kong, China.
- Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2), 249-254.
- Lynn Carlson, John M Conroy, Daniel Marcu, Dianne P O'Leary, Mary E Okurowski, Anthony Taylor, et al. 2001. *An empirical study on the relation between abstracts, extracts, and the discourse structure of texts*. Paper presented at the DUC-2001, New Orleans, LA, USA.
- Lynn Carlson, Daniel Marcu, & Mary E Okurowski. 2002. *RST Discourse Treebank*. Philadelphia, PA: Linguistic Data Consortium.
- Lynn Carlson, Daniel Marcu, & Mary E Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In J. van Kuppevelt & R. Smith (Eds.), *Current directions in discourse and dialogue*. New York: Kluwer Academic Publishers.
- Simon Corston-Oliver. 1998. *Computing representations of the structure of written discourse*. Redmont, WA.
- Chris Ding, Xiaofeng He, Perry Husbands, Hongyuan Zha, & Horst Simon. 2002. *PageRank, HITS, and a unified framework for link analysis*. (No. 49372). Berkeley, CA, USA.
- Jade Goldstein, Mark Kantrowitz, Vibhu O Mittal, & Jamie O Carbonell. 1999. *Summarizing text documents: Sentence selection and evaluation metrics*. Paper presented at the SIGIR-99, Melbourne, Australia.
- Yihong Gong, & Xin Liu. 2001. *Generic text summarization using relevance measure and latent semantic analysis*. Paper presented at the Annual ACM Conference on Research and Development in Information Retrieval, New Orleans, LA, USA.
- Barbara J Grosz, & Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 175-204.
- Julia Hirschberg, & Christine H Nakatani. 1996. *A prosodic analysis of discourse segments in direction-giving monologues*. Paper presented at the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA.
- Jerry R Hobbs. 1985. *On the coherence and structure of discourse*. Stanford, CA.
- D Horn. 1942. A correction for the effect of tied ranks on the value of the rank difference correlation coefficient. *Journal of Educational Psychology*, 33, 686-690.
- Hongyan Jing, Kathleen R McKeown, Regina Barzilay, & Michael Elhadad. 1998. *Summarization evaluation methods: Experiments and analysis*. Paper presented at the AAAI-98 Spring Symposium on Intelligent Text Summarization, Stanford, CA, USA.
- Alex Lascarides, & Nicholas Asher. 1993. Temporal interpretation, discourse relations and common sense entailment. *Linguistics and Philosophy*, 16(5), 437-493.
- Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 159-165.
- William C Mann, & Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3), 243-281.
- Christopher D Manning, & Hinrich Schuetze. 2000. *Foundations of statistical natural language processing*. Cambridge, MA, USA: MIT Press.
- Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. Cambridge, MA: MIT Press.
- Mandar Mitra, Amit Singhal, & Chris Buckley. 1997. *Automatic text summarization by paragraph extraction*. Paper presented at the ACL/EACL-97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain.
- Kenji Ono, Kazuo Sumita, & Seiji Miike. 1994. *Abstract generation based on rhetorical*

- structure extraction*. Paper presented at the COLING-94, Kyoto, Japan.
- Lawrence Page, Sergey Brin, Rajeev Motwani, & Terry Winograd. 1998. *The PageRank citation ranking: Bringing order to the web*. Stanford, CA.
- Dragomir R Radev, Eduard Hovy, & Kathleen R McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4), 399-408.
- Gerard Salton, & Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5), 513-523.
- Karen Sparck-Jones. 1993. What might be in a summary? In G. Knorz, J. Krause & C. Womser-Hacker (Eds.), *Information retrieval 93: Von der Modellierung zur Anwendung* (pp. 9-26). Konstanz: Universitaetsverlag.
- Karen Sparck-Jones, & Tetsuya Sakai. 2001, September 2001. *Generic summaries for indexing in IR*. Paper presented at the ACM SIGIR-2001, New Orleans, LA, USA.
- Klaus Zechner. 1996. *Fast generation of abstracts from general domain text corpora by extracting relevant sentences*. Paper presented at the COLING-96, Copenhagen, Denmark.