Document Fusion for Comprehensive Event Description

Christof Monz
Institute for Logic, Language and Computation
University of Amsterdam
1018 TV Amsterdam, The Netherlands
christof@science.uva.nl
www.science.uva.nl/~christof

Abstract

This paper describes a fully implemented system for fusing related news stories into a single comprehensive description of an event. The basic components and the underlying algorithm are explained. The system uses a computationally feasible and robust notion of entailment for comparing information stemming from different documents. We discuss the issue of evaluating document fusion and provide some preliminary results.

1 Introduction

Conventional text retrieval systems respond to a user's query by providing a (ranked) list of documents which potentially satisfy the information need. After having identified a number of documents which are actually relevant, the user reads some of those documents to get the information requested. To be sure to get a comprehensive account of a particular topic, the list of documents one has to read may be rather long, including a severe amount of redundancy; i.e., documents partially conveying the same information.

Although this problems basically holds for any text retrieval situation, where comprehensiveness is relevant, it becomes particularly evident in the retrieval of news texts.

News agencies, such as AP, BBC, CNN, or Reuters, often describe the same event differently. For instance, they provide different background information, helping the reader to situate the story, they interview different people to comment on an event, and they provide additional, conflicting or more accurate information, depending on their sources.

To get a description of an event which is as comprehensive as possible and also as short as possible, a user has to compile his or her own description by taking parts of the original news stories, ignoring duplicate information. Typical users include journalists and intelligence analysts, for whom compiling and fusing information is an integral part of their work (Carbonell et al., 2000). Obviously, if done manually, this process can be rather laborious as it involves numerous comparisons, depending on the number and length of the documents.

The aim of this paper is to describe an approach automatizing this process by fusing information stemming from different documents to generate a single comprehensive document, containing the information of all original documents without repeating information which is conveyed by two or more documents.

The work described in this paper is closely related to the area of multi-document summarization (Barzilay et al., 1999; Mani and Bloedorn, 1999; McKeown and Radev, 1995; Radev, 2000), where related documents are analyzed to use frequently occurring segments for identifying relevant information that has to be included in the summary. Our work differs from the work on multi-document summarization as we focus on document fusion disregarding summarization. On the contrary, we are not aiming for the shortest description containing the most relevant information, but for the shortest description containing *all* information. For instance, even historic background information is included, as long as it allows the reader to get a more comprehensive description of an event.

Although the techniques that are used for multi-document fusion and multi-document summarization are similar, the task of fusion is complementary to the summarization task. They differ in the way that, roughly speaking, multi-document summarization is the intersection of information within a topic, whereas multi-document fusion is the union of information. They are similar to the extent that in both cases nearly equivalent information stemming from different documents within the topic has to be identified as such.

The remainder of this paper is structured as follows: Section 2 introduces the main components and challenges of implementing a document fusion system. Issues of evaluating document fusion and some preliminary evaluation of our system are presented in Section 3. In Section 4, some conclusions and prospects on future work are given.

2 Fusing Documents

Before developing a document fusion system, some basic issues have to be considered.

1. On which level of granularity are the documents fused (i.e., word or phrase level, sentence level, or paragraph level?
2. How to decide whether news fragments from different sources convey the same information?
3. How to ensure readability of the fused document? I.e., where should information stemming from different documents be placed in the fused document, retaining a natural flow of information.

Each of the these issues is addressed in the following subsections.

2.1 Segmentation

In the current implementation, we decided to fuse documents on paragraph level for two reasons: First, paragraphs are less context-dependent than sentences and are therefore easier to compare. Second, compiling paragraphs yields a better readability of the fused document. It should be noted that paragraphs are rather short in news stories, rarely being longer than three sentences.

When putting together (fusing) pieces of text from different sources in a way that was not anticipated by the writers of the news stories, it can introduce information gaps. For instance, if a paragraph containing a pronoun is taken out of its original context and placed in a new context (the fused document), this can lead to *dangling pronouns*, which cannot be correctly resolved anymore. In general, this problem does not only hold for pronouns but for all kind of anaphoric expressions such as pronouns, definite noun phrases (e.g., *the negotiations*) and anaphoric adverbials (e.g., *later*). To cope with this problem simple segmentation is applied as a pre-processing step where paragraphs that contain pronouns or simple definite noun phrases are attached to the preceding paragraph. A more sophisticated approach to text segmentation is described in (Hearst, 1997).

Obviously, it would be better to use an automatic anaphora resolution component to cope with this problem, see, e.g., (Kennedy and Boguraev, 1996; Kameyama, 1997), where anaphoric expressions are replaced by their antecedents, but at the moment, the integration of such a component remains future work.

2.2 Informativity

(Radev, 2000) describes 24 cross-document relations that can hold between their segments, one of which is the subsumption (or entailment) relation. In the context of document fusion, we focus on the entailment relation and how it can be formally defined; unfortunately, (Radev, 2000) provides no formal definition for any of the relations.

Computing the informativity of a segment compared to another segment is an essential task during document fusion. Here, we say that the i-th segment of document d ($s_{i,d}$) is more informative than the j-th segment of document d' ($s_{j,d'}$) if $s_{i,d}$ entails $s_{j,d'}$. In theory, this should be proven logically, but in practice this is far beyond the current state of the art in natural language processing. Additionally, a binary logical decision might also be too strict for simulating the human understanding of entailment.

A simple but nevertheless quite effective solution is based on one of the simpler similarity measures in information retrieval (IR), where texts are simply represented as bags of (weighted) words.

The definition of the *entailment score (es)* is given in (1). $es(s_{i,d}, s_{j,d'})$ compares the sum of the weights of terms that appear in both segments to the total sum weights of $s_{j,d'}$.

$$es(s_{i,d}, s_{j,d'}) = \frac{\sum_{t_k \in s_{i,d} \cap s_{j,d'}} idf_k}{\sum_{t_k \in s_{j,d'}} idf_k} \quad (1)$$

The weight of a term t_i is its *inverse document frequency* (idf_i), as defined in (2), where N is the number of all segments in the set of related documents (the topic) and n_i is the number of segments in which the term t_i occurs.

$$idf_i = \log\left(\frac{N}{n_i}\right) \quad (2)$$

Terms which occur in many segments (i.e., for which n_i is rather large), such as *the*, *some*, etc., receive a lower *idf*-score than terms that occur only in a few segments. The underlying intuition of the *idf*-score is that terms with a higher *idf*-score are better suited for discriminating the content of a particular segment from the other segments in the topic, or to put it differently, they are more content-bearing. Note, that the logarithm in (2) is only used for dampening the differences.

The entailment score $es(s_{i,d}, s_{j,d'})$ measures how many of the words of the segment $s_{i,d}$ occur in $s_{j,d'}$, and how important those words are. This is obviously a very shallow approach to entailment computation, but nevertheless it proved to be effective, see (Monz and de Rijke, 2001).

2.3 Implementation

In this subsection, we present the general algorithm underlying the implementation, given a set of documents belonging to topic T. The implementation has to tackle two basic tasks. First, identify segments that are entailed by other segments and use the more informative one. Second, place the remaining segments at positions with similar content. The fusion algorithm depicted in Figure 1 consists of five steps.

1. is basically a pre-processing step as explained above. 2. computes pairwise the cross-document *entailment scores* for all segments in T. Although the pairwise computation of es and sim is exponential in the number of documents in T, it still remains computationally tractable in practice. For instance, for a topic containing 4 documents (the average case) it takes 10 CPU seconds to compute all entailment and similarity relations. For a topic containing 8 documents (an artificially constructed extreme case) it takes 66 CPU seconds; both on a 600 MHz Pentium III PC.

In 3., one of the documents is taken as base for the fusion process. Starting with a 'real' document improves the readability of the final fused documents as it imposes some structure on the fusion process. There are several ways to select the base document. For instance, take the document with the most unique terms, or the document with the highest document weight (sum of all *idf*-scores). In the current implementation we simply took the longest document within the topic, which ensures a good base coverage of an event.

4. and 5. are the actual fusion steps. Step 4. replaces a segment s_{i,d_F} in the fused document by a segment $s_{j,d'}$ from another document if $s_{j,d'}$ is the segment maximally entailing s_{i,d_F} and if it is significantly (above the threshold θ_{es}) more informative than s_{i,d_F}. Choosing an optimal value for θ_{es} is essential for the effectiveness of the fusion system. Section 3 discusses some of our experiments to determine θ_{es}.

Step 5. is kind of complementary to step 4., where related but more informative segments are identified. Step 5. identifies segments that add new information to d_F, where a segment $s_{j,d'}$ is new if it has low similarity to all segments in d_F, i.e., if the the similarity score is below the threshold θ_{sim}. If a segment $s_{j,d'}$ is new, it is placed right after the segment in d_F to which it is most similar.

Similarity is implemented as the traditional cosine similarity in information retrieval, as defined in (3). This similarity measure is also known as the *tfc.tfc* measure, see (Salton and Buckley, 1988).

$$sim(s_{i,d}, s_{j,d'}) = \frac{\sum_{t_k \in s_{i,d} \cap s_{j,d'}} w_{k,s_{i,d}} \cdot w_{k,s_{j,d'}}}{\sqrt{\sum_{t_k \in s_{i,d}} w^2_{s_{i,d}} \cdot \sum_{t_k \in s_{j,d'}} w^2_{s_{j,d'}}}} \quad (3)$$

Where $w_{k,s_{i,d}}$ is the weight associated with the term t_k in segment $s_{i,d}$. In the nominator of (3),

1. segmentize all documents in T
2. for all $s_{i,d}, s_{j,d'}$ s.t. $d, d' \in T$ and $d \neq d'$: compute $es(s_{i,d}, s_{j,d'})$
3. select a document $d \in T$ as fusion base document: d_F
4. for all s_{i,d_F}: find $s_{j,d'}$ s.t. $d_F \neq d'$ and $s_{j,d'} = \arg\max_{s_{k,d'}} : \quad es(s_{k,d'}, s_{i,d_F}) > es(s_{i,d_F}, s_{k,d'})$
   if $es(s_{j,d'}, s_{i,d_F}) > \theta_{es}$ then replace s_{i,d_F} by $s_{j,d'}$ in the fused document
5. for all $s_{j,d'}$ s.t. $s_{j,d'} \notin d_F$:
   if for all s_{i,d_F}: $sim(s_{i,d_F}, s_{j,d'}) < \theta_{sim}$,
   then find the most similar s_{i,d_F}: $s_{i,d_F} = \arg\max_{s_{k,d_F}} : \quad sim(s_{j,d'}, s_{k,d_F})$
   and place $s_{j,d'}$ between s_{i,d_F} and s_{i+1,d_F}

Figure 1: Sketch of the document fusion algorithm.

the weights of the terms that occur in $s_{i,d}$ and $s_{j,d'}$ are summed up. The denominator is used for normalization. Otherwise, longer documents tend to result in a higher similarity score. In the current implementation $w_{k,s_{i,d}} = idf_k$ for all $s_{i,d}$. The reader is referred to (Salton and Buckley, 1988; Zobel and Moffat, 1998) for a broad spectrum of similarity measures for information retrieval.

3 Evaluation Issues

The document fusion system is evaluated in two steps. First, the effectiveness of entailment detection is evaluated, which is the key component of our system. Then we present some preliminary evaluation of the whole system focusing on the quality of the fused documents.

3.1 Evaluating Entailment

Recently, we have started to build a small test collection for evaluating entailment relations. The reader is referred to (Monz and de Rijke, 2001) for more details on the results presented in this subsection.

For each of the 21 topics in our test corpus two documents in the topic were randomly selected, and given to a human assessor to determine all subsumption relations between segments in different documents (within the same topic). Judgments were made on a scale 0–2, according to the extent to which one segment was found to entail another.

Out of the 12083 possible subsumption relations between the text segments, 501 (4.15%) received a score of 1, and 89 (0.73%) received a score of 2.

Let a *subsumption pair* be an ordered pair of segments $(s_{i,d}, s_{j,d'})$ that may or may not stand in the subsumption relation, and let a *correct* subsumption pair be a subsumption pair $(s_{i,d}, s_{j,d'})$ for which $s_{i,d}$ does indeed entail $s_{j,d'}$. Further, a *computed* subsumption pair is a subsumption pair for which our subsumption method has produced a score above the subsumption threshold.

Then, precision is the fraction of computed subsumption pairs that is correct:

$$\text{Precision} = \frac{\text{number of correct subsumption pairs computed}}{\text{total number of subsumption pairs computed}}.$$

And recall is the proportion of the total number of correct subsumption pairs that were computed:

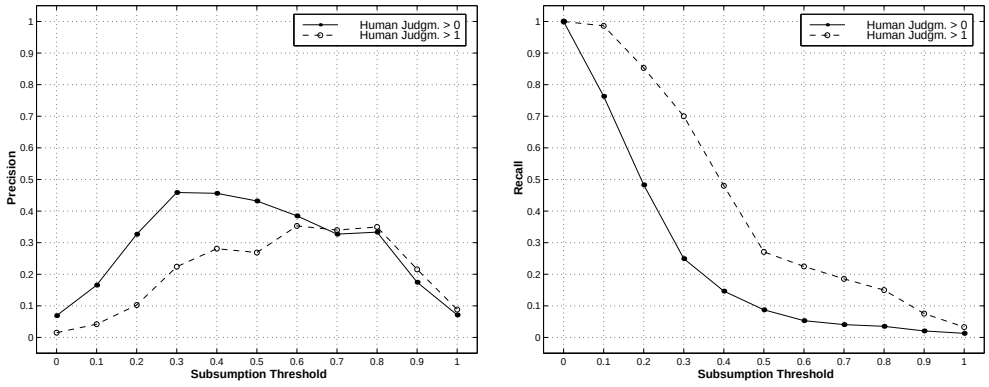
$$\text{Recall} = \frac{\text{number of correct subsumption pairs computed}}{\text{total number of correct subsumption pairs}}.$$

Observe that precision and recall depend on the subsumption threshold that we use.

We computed average recall and precision at 11 different subsumption thresholds, ranging from 0 to 1, with .1 increments; the average was computed over all topics. The results are summarized in Figures 2 (a) and (b).

Since precision and recall suggest two different optimal subsumption thresholds, we use the F-Score, or harmonic mean, which has a high value only when both recall and precision are high.

$$F = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$



(a) (b)

Figure 2: (a) Average precision with human judgments > 0 and > 1. (b) Average recall with human judgments > 0 and > 1.

The average F-scores are given in Figure 3.

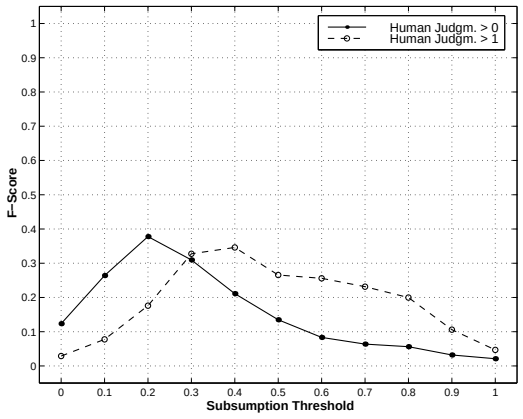


Figure 3: Average F-scores with human judgments > 0 and > 1.

The optimal subsumption threshold for human judgments > 0 is around 0.18, while it is approximately 0.4 for human judgments > 1. This confirms the intuition that a higher threshold is more effective when human judgments are stricter.

3.2 Evaluating Fusion

In the introduction, it was pointed out that document fusion by hand can be rather laborious, and the same holds for the evaluation of automatic document fusion. Similar to automatic summarization, there are no standard document collections or clear evaluation criteria aiding to automatize the process of evaluation. One approach could be to focus on news stories which mention their sources. For instance CNN's new stories often say that "AP and Reuters contributed to this story". On the other hand one has to be cautious to take those news stories as *gold standard* as the respective contributions of the journalist and his or her sources are not made explicit.

In the area of multi-document summarization, there is a distinction between *intrinsic* and *extrinsic* evaluation, see (Mani et al., 1998). Intrinsic evaluation judges the quality directly based on analysis of the summary. Usually, a human judge assesses the quality of a summary based on some standardized evaluation criteria.

In extrinsic evaluation, the usefulness of a summary is judged based on how it affects the completion of some other task. A typical task used for extrinsic evaluation is ad-hoc retrieval, where the relevance of a retrieved document is assessed by a human judge based on the document's summary. Then, those judgments are compared to judgments based on original documents, see, e.g., (Brandow et al., 1995; Mani and Bloedorn, 1999).

At this stage we have just carried out some preliminary evaluation. The test collection consists of 69 news stories categorized into 21 topics. Categorization was done by hand, but it is also possible to have information filtering, see (Robertson and Hull, 2001), or topic detection and tracking (TDT) tools carrying out this task (Allan et al., 1998). All documents belonging to the same topic were released on the same day and describe the same event. Table 1 provides further details on

the collection.

	avg. per topic
no. of docs.	3.3 docs.
length of a doc.	612 words
length of all docs. together	2115 words
length of longest doc.	783 words
length of shortest doc.	444 words

Table 1: Test collection (21 topics, 69 documents).

In addition to the aforementioned news agencies, the collection includes texts from the L.A. Times, Washington Post and Washington Times.

In general, a segment should be included in the fused document if it did not occur before to avoid redundancy (*False Alarm*), and if it adds information, so no information is left out (*Miss*). As in IR or TDT, *Miss* and *False Alarm* tend to be inversely related; i.e., a decrease of *Miss* often results in an increase of *False Alarm* and vice versa.

Table 2 illustrates the different possibilities how the system responds as to whether a segment should be included in the fused document and how a human reader judges.

system judgement	*reader: include*	*reader: exclude*
include	a	b
exclude	c	d

Table 2: Contingency table.

Then, *Miss* and *False Alarm* can be defined as in (4) and (5), respectively.

$$Miss = \frac{c}{a+c} \text{ if } a+c>0 \quad (4)$$

$$False\ Alarm = \frac{b}{b+d} \text{ if } b+d>0 \quad (5)$$

The *fusion impact factor (fif)* describes to what extent the different sources actually contributed to the fused document. For instance if the fused document solely contains segments from one source, *fif* equals 0, and if all sources equally contributed it equals 1. This can be formalized as follows:

$$fif = 1 - \frac{\sum_{d \in T} \mid 1/n_T - n_{seg,d}/n_{seg} \mid}{2 \cdot (1 - 1/n_T)} \quad (6)$$

Where S is a set of related documents, and n_T is its size. n_{seg} is the number of segments in the fused document and $n_{seg,d}$ is the number of segments stemming from document d.

For our test collection, the average *fusion impact factor* was 0.56. Of course the *fif*-score depends on the choice of θ_{es} and θ_{sim}, in a way that a lower value of θ_{es} or a higher value of θ_{sim} increases the *fif*-score. In this case, $\theta_{es} = 0.2$ and $\theta_{sim} = 0.05$.

Table 3 shows the length of the fused documents in average compared to the longest, shortest, and all documents in a topic, for $\theta_{es} = 0.2$ and $\theta_{sim} = 0.05$.

	avg. compression ratio per topic
all docs. together	0.55
longest doc.	1.36
shortest doc.	2.55

Table 3: Compression ratios.

Measuring *Miss* intrinsically is extremely laborious; especially comparing the effectiveness of different values for the thresholds θ_{es} and θ_{sim} is infeasible in practice. Therefore, we decided to measure *Miss* extrinsically. We used ad-hoc retrieval as the extrinsic evaluation task. The evaluation criterion is stated as follows: Using the fused document of each topic as a query, what is the average (non-interpolated) precision?

As baseline, we concatenated all documents of each topic. This would constitute an event description that does not miss any information within the topic. This document is then used to query a collection of 242,996 documents, containing the 69 documents from our test collection. Since the baseline is simply the concatenation of all documents within the topic, one can expect that all documents from that topic receive a high rank in the set of retrieved documents. This average precision forms the optimal performance for that topic. For instance, if a topic contains three documents, and the ad-hoc retrieval ranks those documents as 1, 3, and 6, there are three recall levels: $33.\bar{3}\%$, $66.\bar{6}\%$, and 100%. The precision at these levels is 1/1, 2/3, and 3/6 respectively,

which averages to $0.7\bar{2}$.

The next step is to compare the actually fused documents to the baseline. It is to be expected that the performance is worse, because the fused documents do not contain segments which are entailed by other segments in the topic. For instance, if the fused document for the aforementioned topic is used as a query and the original documents of the topic are ranked as 2, 4, and 9, the average precision is (1/2 + 2/4 + 3/9)/3 = $0.\bar{4}$.

Compared to $0.7\bar{2}$ for the baseline, fusion leads to a decrease of effectiveness of approximately 38.5%. Figure 4, gives the averaged precision for the different values for θ_{es}.

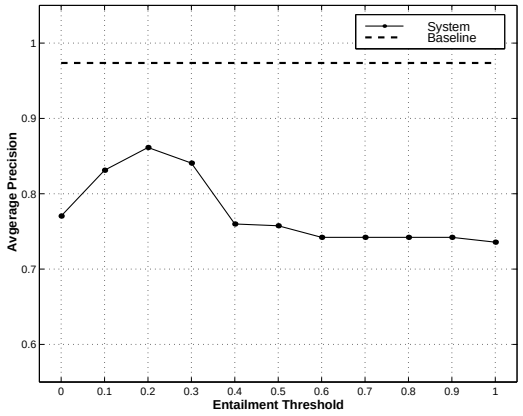


Figure 4: Average precision for ad-hoc retrieval.

It is not obvious how to interpret the numerical value of the ad-hoc retrieval precision in terms of *Miss*, but the degree of deviation from the baseline gives a rough estimate of the completeness of a fused document. At least, this allows for an ordinally scaled ranking of the different methods (in our case different values for θ_{es}), that are used for generating the fused documents. Figure 4 illustrates that in the context of the ad-hoc retrieval evaluation an optimal entailment threshold (θ_{es}) lies around 0.2. Table 4 shows the decrease in retrieval effectiveness in percent, compared to the baseline. The average precision at 0.2 is 0.8614, which is just $\approx 11.5\%$ below the baseline.

For all ad-hoc retrieval experiments, the *Lnu.ltu* weighting scheme, see (Singhal et al., 1996), has been used, which is one of the best-performing weighting schemes in ad-hoc retrieval. In addition to the 69 documents from our collection, the retrieval collection contains articles from Associated Press 1988–1990 (from the TREC distribution), which also belong to the newswire or newspaper domain. Any meta information such as the name of the journalist or news agency is removed to avoid matches based on that information.

θ_{es}	*Decrease in precision*	θ_{es}	*Decrease in precision*
0.0	20.9%	0.6	23.8%
0.1	14.6%	0.7	23.8%
0.2	11.5%	0.8	23.8%
0.3	13.7%	0.9	23.8%
0.4	22.0%	1.0	24.4%
0.5	22.2%		

Table 4: Differences to baseline retrieval.

In the context of multi-document summarization, (Stein et al., 2000) use topic clustering for extrinsic evaluation. Although we did not carry out any evaluation based on topic clustering, it seems that it could also be applied to multi-document fusion, given the close relationship between fusion and summarization on the one hand and retrieval and clustering on the other hand.

4 Conclusions

The document fusion system described is just prototype and there is much more space for improvement. Although detecting redundancies by using a shallow notion of entailment works reasonably well, it is still far from perfect.

In the current implementation, text analysis is very shallow. Pattern matching is used to avoid dangling anaphora and lemmatization is used to make the entailment and similarity scores unsusceptible to morphological variations such as number and tense. A question for future research is to what extent shallow parsing techniques can improve the entailment scores. In particular, does considering the relational structure of a sentence improve computing entailment relations? This has shown to be successful in inference-based approaches to question-answering, see (Harabagiu et al., 2000), and document fusion might also benefit from representations that are a bit deeper than the one discussed in this paper.

Another open issue at this point is the need for standards for evaluating the quality of document

fusion. We think that this can be done by using standard IR measures like *Miss* and *False Alarm*. Although *Miss* can be approximated extrinsically, it is unclear whether this also possible for *False Alarm*. Obviously, intrinsic evaluation is more reliable, but it remains an extremely laborious process, where inter-judge disagreement is still an issue, see (Radev et al., 2000).

Acknowledgments

The author would like to thank Maarten de Rijke for providing the entailment judgments. This work was supported by the Physical Sciences Council with financial support from the Netherlands Organization for Scientific Research (NWO), project 612-13-001.

References

J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. 1998. Topic detection and tracking pilot study final report. In *Proceedings of the Broadcast News Transcription and Understranding Workshop (Sponsored by DARPA)*.

R. Barzilay, K. McKeown, and M. Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL'99)*.

R. Brandow, K. Mitze, and L. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, 31(5):675–685.

J. Carbonell, D. Harman, E. Hovy, S. Maiorano, J. Prange, and Sparck-Jones. K. 2000. Vision statement to guide research in question answering (Q&A) and text summarization. NIST Draft Publication.

S. Harabagiu, M. Pasca, and S. Maiorano. 2000. Experiments with open-domain textual question answering. In *Proceedings of COLING-2000*.

M. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

M. Kameyama. 1997. Recognizing referential links: An information extraction perspective. In R. Mitkov and B. Boguraev, editors, *Proceedings of ACL/EACL-97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 46–53.

C. Kennedy and B. Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*. Association for Computational Linguistics.

I. Mani and E. Bloedorn. 1999. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1–2):35–67.

I. Mani, D. House, G. Klein, L. Hirschman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. 1998. The Tipster SUMMAC text summarization evaluation, final report. Technical Report 98W0000138, Mitre.

K. McKeown and D. Radev. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82.

C. Monz and M. de Rijke. 2001. Light-weight subsumption checking for computational semantics. In P. Blackburn and M. Kohlhase, editors, *Proceedings of the 3rd Workshop on Inference in Computational Semantics (ICoS-3)*.

D. Radev, H. Jing, and M. Budzikowska. 2000. Centroid-based summarization of multiple documents: Clustering, sentence extraction, and evaluation. In *Proceedings of the ANLP/NAACL-2000 Workshop on Summarization*.

D. Radev. 2000. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In *In Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*.

S. Robertson and D. Hull. 2001. The TREC-9 filtering track final report. In *Proceedings of The 9th Text Retrieval Conference (TREC-9)*. NIST Special Publication.

G. Salton and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

A. Singhal, G. Salton, M. Mitra, and C. Buckley. 1996. Document length normalization. *Information Processing & Management*, 32(5):619–633.

G. Stein, G. Wise, T. Strzalkowski, and A. Bagga. 2000. Evaluating summaries for multiple documents in an interactive environment. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, pages 1651–1657.

J. Zobel and A. Moffat. 1998. Exploring the similarity space. *ACM SIGIR Forum*, 32(1):18–34.