Using Document Features and Statistical Modeling to Improve Query-Based Summarization

Judith D. Schlesinger
IDA/Center for Computing Sciences
judith@super.org

Deborah J. Baker
Department of Defense
debbaker@hotmail.com

Robert L. Donaway
Logicon Technology Solutions
Knowledge Services Division
rdonaway@logicon.com

Abstract

We report on an effort to improve (in terms of both higher precision and recall values *and* more acceptable to system users) the indicative query-based summaries generated by an existing *operational* text extraction system (document summarizer) that produces both generic and query-based summaries. An extensive analysis of sentence- and document-related features coupled with a logistic-regression model have yielded summaries that produce f-scores that range from 43–67% higher than those of the original algorithm.

1 Introduction

There are many different ways to classify automatic summarization systems, including extraction vs. abstraction, knowledge-rich (or intensive) vs. knowledge-poor (or scant), and indicative vs. informative. While either an indicative or an informative summary can be generated for any of the other classifications, text abstraction relies on knowledge-rich approaches while text extraction can be done with either knowledge-rich or knowledge-poor methods ([HM 00]). Knowledge-rich approaches strive for an understanding of the text to be summarized, using in-depth parsing, frames, and discourse analysis (which may also be used for knowledge-poor approaches), in order to generate a coherent, effective summary.

The summarization system we are discussing here is an indicative extraction system that uses a knowledge-poor, mathematical modeling approach. Therefore, we limit further discussion to this type of summarizer.

Text extraction relies on sentence ([P 90], [BMR95]) or paragraph ([SAB94], [MSB97]) extraction. The main thrust of extraction is to select a few "representative" passages from the source document which convey the content either generally, or specifically relating to query terms. This selection is done using either knowledge-rich methods that usually require the ability to "fill" a pre-determined template and, therefore, are domain specific ([GS 93], [JKR93], [F 98], [MJH98]), or knowledge-poor methods which are typically accomplished by developing a way to score the sentences (or paragraphs) of the document and selecting those with the best score, with or without reordering. A brief list of some of the recent work using knowledge-poor methods includes [KPC95], [AL 97], [TM 97], [CG 98], [HL 99], [MJ 99], [BM 00], and, of course, our original system (see Section 2). The method used to generate a score is the typical distinguishing feature amongst these summarizers.

A major problem with this approach is that the results may be unnatural because there is no guarantee that the selected sentences form a coherent, cohesive summary but the approach is used nonetheless because of the *relative* success of the technique.

Most of the current effort involves generic summarization. User supplied query terms have been used in only a very limited number of the summarizers in the literature, including [KPC95] and [TS 98], although they are common in general information retrieval work.

In [KPC95], Kupiec et al developed a trainable summarization program using a combination of "cue words" (words appearing often in summary sentences), keywords, and the position of a sentence in the document. Given a training set of documents

with hand-selected document extracts, a classification function that estimates the probability that a given sentence is included in an extract was developed. [TS 98] uses document summarization "biased" by the query terms supplied by the user as a way to assist in evaluating and improving an information retrieval system. Summary sentences are extracted by calculating a score for each sentence using features including the document title, the location of a sentence in the document, clusters of significant words, and the occurrence of user supplied query terms.

In [MJ 99], Myaeng and Jang use an approach similar to Kupiec et al ([KPC95]) for Korean texts. In addition to the features listed for [KPC95], a measure of similarity between a sentence and the rest of the document, a measure of similarity between a sentence and the document title, and a text component feature that divides a document into several parts and determines if a sentence belongs to the "major content" component have been added. This last can be used as either a feature or a filter. The features in [MJ 99] are all computed independently and then combined using the Dempster-Shafer combination rule.

The operational system combines the approaches of both [KPC95] and [MJ 99]. While features and a statistical model are used as with [KPC95], our system also uses the tf*idf score which is based on IR-related word frequencies. Our work, while related to the original system, of course, is more representative of [KPC95] since we do not require any corpus-based statistics, such as term frequency counts, a priori. We use a logistic regression model, where the features are *not* treated independently from one another. We also use a different set of features from all of these. For example, due to the wide variety in our data, we cannot assume that documents have titles. Also, since we are looking at query-based summaries, a feature such as the major content component is not needed.

The system we developed consists of two parts, a training system and a summary generator. Both utilize the same set of procedures to gather feature information on each sentence in a document. The training system runs the features through a logistic regression model to generate the required coefficients for the model; the summary generator uses the generated coefficients and the accumulated feature information to score each sentence. A summary is created by selecting some number of highest scoring sentences.

Our system operates on documents contained in a highly heterogeneous proprietary corpus of news articles. We also have a substantial data base of TREC documents which is linked to the system as well. By requiring our algorithm to perform well on two dissimilar corpora, we expect to be able to generalize to many different types of documents.

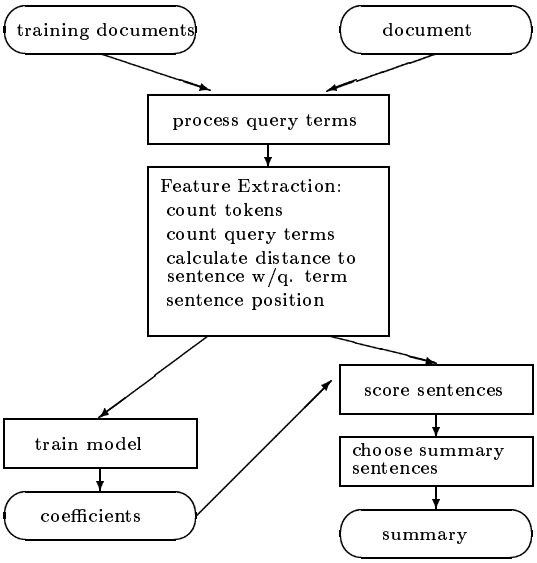

Figure 1: Overview of the Summary System

Figure 1 shows an overview of our method, which is described in detail in Section 3.

2 The Original System

The operational system has been described in some detail in [AOG97], [AOG98], and [AGL99]. We briefly describe it, focusing on the query-based summarization algorithm and related issues.

The summarization system is a frequency-based sentence extractor that applies proven NLP techniques such as corpus-based statistical NLP, robust information extraction, and readily available on-line resources such as WordNet ([Fe 98]) and the Brill POS tagger ([Br 93]). The system uses linguistic motivation and an automatically derived set of features to calculate sentence worthiness. The system is trained using a Bayesian model, incorporating the best conceptual feature set capturing *signature terms* and conventional positional information. We refer the reader to the cited papers for the motivation and theory supporting this.

The system uses four features—(1) short sentence length (< 5 words), (2) sentence position in the document (defined by quadrants), (3) sentence

position in the paragraph (first, middle, last), and (4) inclusion of high tf*idf signature words in a sentence—to select summary sentences. For query-based summaries, this last becomes the tf*idf score of each unique query term in a sentence multiplied by 100. This effectively dominates the scoring algorithm so the algorithm *only* chooses sentences which contain query terms unless none exist (which is unlikely in our user environment[1]).

It is a common weakness of frequency-based systems to generate summaries that are disconnected and difficult for the user to read. Our system's tendency to choose only those sentences that contain query-terms exacerbates this problem.

3 The New Query-Based Summarization Model

As mentioned earlier, we worked with two corpora with a combined tagged set of approximately 1600 documents. The first is a proprietary corpus consisting of news articles of various sizes and formats. Due to the variety of styles, neither the original summarizing algorithm nor any new algorithm, could rely on any "clues" derived from the document format.

For this corpora, we had a set of approximately 700 documents that were tagged by one or more readers. This set was divided into two roughly equal parts, one used for training and the other for testing.

Our second corpus is comprised of documents in 18 TREC topics[2] for a total of just over 900 documents. Topics 110, 132, and 138 were tagged by two or more readers and were used as the training set, while all the others were tagged by just a single person and used only for testing.

3.1 Feature Evaluation

We began by brainstorming a list of features that could be extracted from a (tagged) document that we thought would be useful for summary generation. Figure 2 contains a complete list of the features we considered.

We then culled this list. Feature 12 was eliminated since we could not find a way to use the information in the statistical models with which we were working. Features 13 and 14 were relegated to future efforts

1. position of summary sentence in document
2. percentage (based on number of sentences) of a document contained in a summary
3. position of summary sentence in paragraph—first, middle, last
4. number of paragraph (from 1 to N) from which summary sentence was chosen
5. number of unique query terms in a summary sentence
6. frequency of query terms, i.e., the non-unique count, in a summary sentence
7. density of query terms in a summary sentence
8. number of tokens (non-stop words) in a summary sentence
9. distance of a summary sentence from a sentence (summary or not) that contains a query term; distance = 0 if sentence contains a query term
10. query-term containing sentences *not* in the summary
11. identification of "core" and "background" summary sentences
12. relationship(s) between multiple query terms in a summary sentence—adjacent, within an x-word window, independent, etc.
13. collocations that occur in summary sentence text
14. discourse markers that occur in summary sentences

Figure 2: Initial List of Features

(see Section 5). Other features were eliminated as we experimented.

Analysis of the human-generated (extraction) summaries yielded two interesting examples of known coherence relationships. First, approximately one-third of the tagged summaries contained one or more sentences with *no* query terms. These sentences either (a) contribute essential background/elaboration information or (b) serve the purpose of creating cohesion between selected sentences which contain query term(s), i.e., they help the "flow" of the summary, or (c) both. Therefore, it is clear that something more than the existence of a query term in a sentence is needed for good sentence selection.

Second, the concept of core and background sentences evolved. It was common, but not universal, for the different readers tagging the same document to choose one or more of the same sentences for their summary. "Core" sentences are those chosen by *all* taggers. Any sentences chosen by one or more, but not all, of the readers are called "background" sentences. The position of these background sentences varied but were most often clustered about a core sentence. Often, it was the background sentences that did not contain any query terms, and, while the taggers agreed that these background sentences were needed, they

[1] If no sentences can be found, the system resorts to the generic summary algorithm, using the other features.

[2] Topic numbers 110, 127, 132, 138, 141, 151, 162, 182, 198, 200, 257, 269, 273, 274, 285, 286, 288, and 297.

V1. Number of unique query terms in a sentence.

V2. Number of tokens in a sentence

V3. The distance of a sentence from one with a query term

V4. The position of a sentence in a document

V5. Paragraph position (start, middle, or end)

Figure 3: List of the Most Predictive Features

didn't necessarily agree on which sentence(s) to select.

Much of our effort with feature extraction was directed at trying to identify these non-query term, background sentences.

3.2 Model Development

After collecting all the information about the documents and summaries in our training data sets, we modeled the sentence extraction patterns using a simple linear regression with responses 0, 1, and 2 for non-extracted, background, and core, respectively, and found that certain features were non-predictive. These included the paragraph number, the frequency count of query terms in the sentence, and the density of query terms in the sentence (the ratio of query terms to total length), probably since they are redundant to the information supplied by the unique query terms count.

We made several observations about the features the model deemed most predictive (shown in Figure 3). We found that the more unique query terms a sentence has, the more important it is. Longer sentences tend to be more important. If a sentence does not contain a query term, the closer it is to a query term containing sentence, the more likely it is to belong in the summary. We also found that there is a bias for summary sentences to come from earlier in the document.

There appears to be two reasons for this: (1) there is a natural human inclination to stop tagging summary sentences once a reasonable set has been selected even if the best sentences have not yet been tagged; and (2) if a document's focus matches that of the query terms, it is the writer's natural tendency to place sentences that are useful to a summary early in the document. While the second of these creates a legitimate bias towards earlier sentences, the first does not.

We also found that it is difficult to evaluate summaries solely by comparing model selected sentence numbers and human selected sentence numbers. We found instances where the machine selected an almost identical sentence to a human selected sentence, only it appeared later in the document, and thus "didn't match". Dealing with this evaluation issue, however, is beyond the scope of this paper.

Our initial evaluation showed that when sentences that the model scored above 1 were labeled "extracted", roughly 50% of the sentences were misclassified (either false negative or false positive). Clearly, we wanted a higher success rate.

We modified (and simplified) the problem by changing our response variable to 0 and 1 for non-extracted and extracted, respectively. In other words, background and core were treated as a single group of extracted sentences. While we were concerned about the loss of information this might incur, we felt it was worth trying since we had not yet developed a model which performed at a level at which we were satisfied.

This reformulation of the problem suggested the logistic regression model, which is designed for a binary response variable. Unlike ordinary linear regression, logistic regression constrains the fitted values to lie between 0 and 1. We also noted that we could improve the model by converting our predictive features to also lie within the 0–1 range. Features V1 and V2, from Figure 3, were both changed by dividing the calculated value by the largest value of that feature appearing in the document, i.e., the greatest number of unique query terms occurring in any sentence in the document and the number of tokens in the longest sentence in the document, respectively. V4 was first changed from a quadrant value to the actual sentence number. It was then divided by the number of sentences in the document. V3, the distance from a sentence with a query term, i.e., 0 if the sentence contains a query term, is a skewed non-negative feature. Moreover its effect on the model *should* decrease as it gets larger. For example, a distance of one sentence should be treated very differently than a distance of 5, but a distance of 10 should only be slightly different than a distance of 5. For these reasons, we transformed V3 using the logarithm: log(1 + V3). During the course of this experimentation, we noticed that V5, paragraph position, was not contributing much information. We eliminated it from the model, leaving just four features.

When this modified model was applied to the training data, using 0.5 as the cut-off, the misclassification rate (again, both false negative and false positive combined) dropped to 25% which was far more acceptable.

The logistic regression formula is as follows:

$$score = f(\alpha + \beta_1(V1) + \beta_2(V2) + \beta_3 \cdot g(V3) + \beta_4(V4))$$

where $f(x) = \frac{e^x}{1+e^x}$ and $g(x) = log(1+x)$

and α and the β_is are defined from the training.

However, this application of the score is not the way it would usually be used. When a summary of a document is generated, all sentences are scored and the top N scoring sentences are selected for the summary. N can be chosen by the user, can be some function of the length of the document, or can be some arbitrary system selected value. The way N is chosen has a significant impact on the results, as can be seen in the next section.

4 Results

As shown below, the new logistic regression model generates significantly different summaries from those generated by the original system. As an example, Figure 4 shows a single TREC document while Figures 5 and 6 show the summary sentences selected by the original system and the logistic regression algorithm, respectively. We should note that the original system was tuned to generate the best possible summaries it could. Purposely minimizing the system's capabilities would serve us no useful benefit.

We ran several experiments which are described below. Our evaluation method is very straightforward. We calculated the f-scores[3] for the summaries generated by the logistic regression model for all of our tagged data as well as the f-scores for all of the summaries generated by the original system on the same data as well as for one tagger's summaries compared to another. The original system scores are considered the lower bounds and the human-human scores are considered the upper bounds[4]. These bounds values are shown in Table 1.

	proprietary data set	TREC data set
original system	.31	.30
human vs. human	.69	.65

Table 1: Upper and Lower Bounds for Evaluation Purposes

[3]Our f-score formula is: $\frac{2*precision*recall}{precision+recall}$.

[4]We realize that it is possible to beat these human-human scores but, nonetheless, are considering them to be our target goal.

International: Mandela says Army of ANC May Be Needed

——

At Party Meeting, He Urges Other Nations to Retain South African Sanctions

——

By Joe Davidson, Staff Reporter of The Wall Street Journal

DURBAN, South Africa – Nelson Mandela accused Pretoria of "pursuing a double agenda" and said the African National Congress army would be ready in case a peaceful road to democracy is blocked.

During his speech opening the ANC's national convention yesterday, some of the loudest applause came when he said the ANC army "has a responsibility to keep itself in a state of readiness in case the forces of counterrevolution once more block the path to a peaceful transition to a democratic society."

Setting a tough tune at the organization's first full conference inside the country in three decades, ANC Deputy President Mandela also told the delegates the organization needs to convince the world not to relax sanctions against South Africa, so as not to "lose this weapon which we will need until a democratic constitution has been adopted."

Many of the 2,000 delegates gathered in a university field house in Durban believe the ANC has given up too much during its pro-democracy talks with the South African government. Even members of the ANC's executive committee have expressed worries that excessive faith may have been placed in South African President F.W. de Klerk and negotiations.

Mr. de Klerk recently called on the ANC to terminate, not just suspend as it has, armed actions against the apartheid government. Mr Mandela rejected that suggestion. In fact, he added that Umkhonto We Sizwe, the ANC army, should "make its expertise available" to violence-plagued black communities setting up self-defense units.

Mr. Mandela, who once called Mr. de Klerk a "man of integrity," accused the de Klerk government "of talking peace while actually conducting war." The ANC doubts Pretoria's "good faith when it sits paralyzed as the security forces it controls themselves engage in violence against the people, permit such violence to occur and remain immune from prosecution when there is clear evidences of their involvement or connivance at the murder of innocent people," he said.

Mr. Mandela's call for continued sanctions comes as the South African media bring almost daily reports of the government's growing international acceptance. ANC officials acknowledge they haven't done a good enough job spreading the word that apartheid practices continue even as the government terminates some, but not all, race-based laws.

U.S. Rep. Maxine Waters, a California Democrat who is part of a small U.S. delegation here, told the gathering the Congressional Black Caucus believes the South African government hasn't done enough to justify the removal of U.S. sanctions against Pretoria. The caucus lobbied against their removal at a June 25 meeting with President Bush. In an interview, Ms. Waters said, "The president is anxious to relax sanctions.

Figure 4: TREC Article—110/WSJ910703-0148

International: Mandela says Army of ANC May Be Needed
During his speech opening the ANC's national convention yesterday, some of the loudest applause came when he said the ANC army "has a responsibility to keep itself in a state of readiness in case the forces of counterrevolution once more block the path to a peaceful transition to a democratic society." In fact, he added that Umkhonto We Sizwe, the ANC army, should "make its expertise available" to violence-plagued black communities setting up self-defense units.

Figure 5: Original System Generated Summary

International: Mandela says Army of ANC May Be Needed
DURBAN, South Africa – Nelson Mandela accused Pretoria of "pursuing a double agenda" and said the African National Congress army would be ready in case a peaceful road to democracy is blocked. Setting a tough tone at the organization's first full conference inside the country in three decades, ANC Deputy President Mandela also told the delegates the organization needs to convince the world not to relax sanctions against South Africa, so as not to "lose this weapon which we will need until a democratic constitution has been adopted."

Figure 6: Logistic-Regression Generated Summary

We trained with two different training sets, the proprietary set and the TREC set, as previously describe, in three different modes: (1) multiple human summaries for the same document were merged by taking a union to form a single set of extracted sentences; (2) each human summary for a document was considered separately; and (3) only one human's summaries was used. Each training set produced a different set of coefficients. Summaries were then generated for our test sets using each set of coefficients and were scored in two ways: first against all the summaries in the test set and then against only summaries of the same length (so recall = precision = f-score). Results are shown in Table 2.

Our results are consistently better than those of the original system, with improvement ranging from a low of 43% to a high of 67%. We've loosely split the difference between our lower and upper bounds, giving us a marked improvement while leaving us plenty of room for additional improvement.

There are some unexplained anomalies in our results. For one, the best result of .5 for the proprietary data set comes from training on a single tagger *where that tagger did not tag any summaries for the corresponding test data set* and the .44 score for the corresponding TREC experiment was on data where the tagger marked *every* document in both the training and test sets.

A second is that the TREC scores were almost always higher when training was done with the proprietary data set rather than the TREC data. A third is the consistently higher scores the proprietary summaries received, as compared with those for the TREC data, even when the TREC data was used for training. This is especially surprising since the proprietary data is far more heterogeneous in nature than the TREC data.

Experiment	proprietary data set	TREC data set
TREC training		
merged—all data	.49	.44
merged—same length	.47	.5
separate—all data	.48	.43
separate—same length	.46	.47
1 tagger—all data	.5	.44
1 tagger—same length	.48	.49
proprietary training		
merged—all data	.48	.45
merged—same length	.49	.48
separate—all data	.45	.45
separate—same length	.45	.48
1 tagger—all data	.48	.45
1 tagger—same length	.49	.48

Table 2: F-scores for Different Training Modes

Clearly, our model is *not* capturing all tagger nuances and corpus patterns. This can be both good—train once and then use on any data set—and bad—missing valuable information that can be used to improve summary quality—and the effects need to be further studied.

5 Future Efforts

We have a lot of work left to do. We first need to evaluate our model in lieu of our results and identify why we're getting some of the unusual results that we are and better understand the strengths and weaknesses of the current model. Following that, we need to find a way to eliminate sentences that score high yet are clearly (to the human) not relevant either because the document covers more than one subject and they are not all on topic or the document is entirely off topic. Related to that, we need to identify a way to decide to make a summary shorter than requested because there aren't enough relevant sentences. We are also continuing to improve the method by which we select the background, i.e., linking, sentences. We have tried several methods that we did not include in this discussion because the results remain unsatisfactory.

Additionally, we want to try some other models,

especially a Hidden Markov Model. An HMM was used by another team in our research group to improve the generic summarization algorithm ([CO 01]). System maintenance would be that much easier if we could use the same model for both types of summaries.

In the long term, we want to add in features that can be derived by discourse analysis, including the height in the tree (important nodes are near the root), the salience of the node (essential vs. supporting), and the relation assigned to the node. Significant work was done with discourse analysis as part of our overall research effort ([COM01]) and we will be linking our work with that. We expect the discourse analysis to assist with coreferencing, the use of collocations that occur in summary sentence text, and phrase extraction, all of which should assist in creating more cohesive summaries.

We ultimately want to address the multi-document and multi-language problems so we expect to be kept busy for some time to come.

References

[AL 97] J. Abracos and G.P. Lopez, "Statistical Methods for Retrieving Most Significant Paragraphs in Newspaper Articles", Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 1997, pp. 51–57.

[AOG97] C. Aone, M.E. Okurowski, J. Gorlinsky, "A Scalable Summarization System Using Robust NLP", Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 1997, pp. 66–73.

[AOG98] C. Aone, M.E. Okurowski, and J. Gorlinsky, "Trainable, Scalable Summarization Using Robust NLP and Machine Learning", Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics, 1998, pp. 62–66.

[AGL99] C. Aone, J. Gorlinsky, B. Larsen, and M.E. Okurowski, "A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques", in I. Mani and M.T. Maybury (eds.), *Advances in Automatic Text Summarization*, The MIT Press, 1999, Chapter 7.

[BM 00] A.L. Berger and V.O. Mittal, "OCELOT: A System for Summarizing Web Pages", Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000, pp. 144–151.

[BMR95] R. Brandow, K. Mitze, and L.F. Rau, "Automatic Condensation of Electronic Publications by Sentence Selection", Information Processing and Management, 31(5), 1995, pp. 675–685.

[Br 93] E. Brill, "A Corpus-based Approach to Language Learning", Ph.D. Dissertation, University of Pennsylvania, 1993.

[CG 98] J.G. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries", Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 335-336.

[COM01] L.M. Carlson, Mary Ellen Okurowski, and D. Marcu, "Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory", submitted to the 39th Annual Meeting of the Association for Computational Linguistics, 2001.

[CO 01] J.M. Conroy and D. P. O'Leary, "Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition", submitted to ACM Transaction on Information Systems.

[Fe 98] C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[F 98] F. Freitag, "Toward General-Purpose Learning for Information Extraction", Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 1998, pp. 404–408.

[GS 93] R. Grishman and J. Sterling, "Description of the Proteus System as Used for MUC-5", Proceedings of MUC-5, 1993, pp. 181–194.

[HM 00] U. Hahn and I. Mani, "The Challenges of Automatic Summarization", Computer, vol. 33, no. 11, November, 2000, pp. 29–36.

[HL 99] E.H. Hovy and C.-Y. Lin, "Automating Text Summarization in SUMMARIST", I. Mani and M.T. Maybury (eds.), *Advances in Automatic Text Summarization*, MIT Press, 1999, Chapter 8.

[JKR93] P.S. Jacobs, G. Krupka, L. Rau, M. Maulden, T. Mitamura, T. Kitani, I. Sider, and L. Childs, "Description of the SHOGUN System Used for MUC-5", Proceedings of MUC-5, 1993, pp.109–120.

[KPC95] J. Kupiec, J. Pederson, and F. Chen, "A Trainable Document Summarizer", Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995, pp. 68–73.

[MJH98] K.R. McKeown, D.A. Jordan, and V. Hatzivassiloglou, "Generating Patient-Specific Summaries of Online Literature", Proceedings of AAAI 98 Spring Symposium on Intelligent Text Summarization, 1998, pp. 34–43.

[MSB97] M. Mitra, A. Singhal, and C. Buckley, "Automatic Text Summarization by Paragraph Extraction", Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 1997, pp. 39–46.

[MJ 99] S.H. Myaeng and D.-H. Jang, "Development and Evaluation of a Statistically-Based Document Summarization System", I. Mani and M.T. Maybury, (eds.), *Advances in Automatic Text Summarization*, The MIT Press, 1999, Chapter 6.

[P 90] C.D. Paice, "Constructing Literature Abstracts by Computer: Techniques and Perspectives", Information Processing and Management, 26(1), 1990, pp. 171–186.

[SAB94] G. Salton, J. Allan, and C. Buckley, and A. Singhal, "Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts", Science, 264, 1994, pp. 1421–1426.

[TM 97] S. Teufel and M. Moens, "Sentence Extraction as a Classification Task", Proceedings of the ACL/EACL Workshop on Intelligent Scalable Text Summarization, 1997, pp. 58–65.

[TS 98] A. Tombros and M. Sanderson, "Advantages of Query Biased Summaries in Information Retrieval", Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 2–10.