Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment

Horacio Saggion[*], Dragomir Radev[**], Simone Teufel[***], Wai Lam[****], Stephanie M. Strassel[*****]

[*]Department of Computer Science
University of Sheffield
211 Portobello Street, Sheffield S1 4DP, England, UK
saggion@dcs.shef.ac.uk

[**]School of Information & Department of Electrical Engineering and Computer Science
University of Michigan
550 E. University
3080 West Hall
Ann Arbor, MI 48109-1092
radev@si.umich.edu

[***]Computer Laboratory
Cambridge University
JJ Thomson Avenue
Cambridge CB3 0FD, UK
Simone.Teufel@cl.cam.ac.uk

[****]Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong
Shatin
Hong Kong
wlam@se.cuhk.edu.hk

[*****]Linguistic Data Consortium
University of Pennsylvania
3615 Market Street
Philadelphia, PA 19104
strassel@ldc.upenn.edu

Abstract

We describe our work on the development of Language and Evaluation Resources for the evaluation of summaries in English and Chinese. The language resources include a parallel corpus of English and Chinese texts which are translations of each other, a set of queries in both languages, clusters of documents relevants to each query, sentence relevance measures for each sentence in the document clusters, and manual multi-document summaries at different compression rates. The evaluation resources consist of metrics for measuring the content of automatic summaries against reference summaries. The framework can be used in the evaluation of extractive, non-extractive, single and multi-document summarization. We focus on the resources developed that are made available for the research community.

1. Introduction

Evaluation is an essential step of any natural language processing task. In the field of text summarization almost all research is published with an in-house evaluation, which makes it difficult to replicate experiments, to compare results, or to use evaluation data for training purposes. The development of standards of evaluation and sharable resources is of paramount importance for progress in text summarization. SUMMAC (Mani et al., 1998) and DUC (2000) are clear examples of efforts to advance text summarization research.

This paper describes the language resources developed for the evaluation of text summarization systems in a cross-lingual environment. These resources have been constructed in the context of the 2001 Workshop on Automatic Summarization of Multiple (Multilingual) Doc-

uments, a 6-week language engineering workshop at the Center for Language and Speech Processing, Johns Hopkins University. The objectives of the workshop were the integration of cross-lingual information retrieval with single and multi-document summarization and its evaluation.

2. Corpus and Annotation

We use a parallel corpus of English and Chinese (Cantonese) texts which are translations or near translations of each other. The corpus consists of 18,461 document-pairs. The corpus, called the *Hong Kong Newspaper Corpus* (corpus number LDC2000T46), is provided by the Linguistic Data Consortium (LDC). The average size in words for a document is 347.8 for English and 325.2 for Chinese, in sentences it is 16.2 and 15.5, respectively. The texts are not typical news articles. The Hong Kong Newspaper mainly publishes announcements of the local administration and descriptions of municipal events, such as an anniversary of the fire department, or seasonal festivals.

Each document in the corpus was annotated in order to provide structural and linguistic information. The annotation for each document includes information about the document identity, its language and its translation. For the purpose of text summarization (or sentence extraction) research it was identified that annotations on the sentence level were required in order to allow fair comparison between different summarization technologies. We provided mark-up on the paragraph, sentence and word level. As one of the research objectives of the workshop was to investigate new measures for content evaluation based on the notion of vocabulary overlap, English documents were also annotated with parts of speech and morphologic information.

2.1. Corpus Processing and Encoding

All corpus information is encoded in XML, and several DTDs were written to describe the structure of the document after each processing step. This proved an advantage for Software engineering in the project, as many modules had to be interfaced, and XML validation made it very easy to check for errors in the input and output of each module in the pipeline.

2.1.1. Processing of English Documents

We automatically separated the main title from the main body of text of the news article, inserted sentence and word boundaries using the LT TTT Tokenisation Tool (Grover et al., 2000), a software package developed within an XML processing paradigm which provides tools for text tokenisation and mark-up. Semi-automatic corrections of sentence boundaries were made in those sets of documents where human sentence segmentation was available. In Figure 1, we show a short document from the corpus annotated with sentence boundaries.

The English corpus was further annotated with part of speech tags[1] (Mikheev, 2000). The tagger is based on a combination of Hidden Markov Models and Maximum Entropy technologies, it comes trained on 4 million words of the Wall Street Journal. A lemmatiser (Humphreys et al., 2000) was also used to obtain lemmas for each noun and verb in the text. It is a rule-based algorithm that produces for each noun and verb in the input text an affix and root. The program is implemented as a set of regular expressions which represent both morphological analyses of the input and exception rules. The program can be considered domain independent because exception rules were derived from WordNet (Miller, 1995), but also revised by the analysis of a number of English corpora. A sentence annotated with parts of speech and lemmas can be seen in Figure 2.

2.1.2. Processing of Chinese Documents

Sentence segmentation in Chinese is based on punctuation. A list of punctuation symbols that usually indicate end of sentences was created. This list was used in conjunction with a greedy matching algorithm over sequences of punctuation symbols as the basis for sentence end identification. Further, we made use of a word segmentation program derived from the tool provided in http://www.mandarintools.com for the BIG5 encoding. Words in Chinese were identified by means of a dictionary used in conjunction with a maximal matching algorithm that attempts to match the longest possible word in the dictionary. The algorithm is able to identify dates, times, person names, locations, money amounts, organization names, and percentages. In Figure 3, we show a small Chinese document annotated with sentence boundaries.

2.1.3. Named Entity Recognition

Named Entity (NE) detection is the process of identifying and categorising names in texts (person, organization, location, date, time, money, and percent). Both Chinese and English text were annotated with named entity tags using *IdentiFinder* (BBN, 2000), a probabilistic natural language software tool that scans text to locate NEs. The tool analyzes training data, counts and compiles statistics about the training data, convert those statistics into probabilistic models, applies

[1] From the Penn Tree-bank tag-set.

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE DOCSENT SYSTEM "/export/ws01summ/dtd/docsent.dtd" >
<DOCSENT DID='D-19980303_004.e' DOCNO='2203' LANG='ENG' CORR-DOC='D-
19980303_004.c'>
<BODY>
<HEADLINE><S PAR="1" RSNT="1" SNO="1"> Joseph W P Wong accepts ATV's apology
</S></HEADLINE>
<TEXT>
<S PAR='2' RSNT='1' SNO='2'>The Secretary for Education and Manpower, Mr Joseph
W P Wong, said today (Tuesday) that he had accepted the apology of Asia Televi-
sion Limited (ATV) over the remarks made on him in the ATV programme "Hong Kong
Affairs" last Monday (February 23) and would not pursue the matter further.</S>
</TEXT>
</BODY>
</DOCSENT>
```

Figure 1: Document 19980303_004.e annotated with sentence boundaries.

```
<S PAR='1' RSNT='1' SNO='1'><W C='NNP' L='joseph'>Joseph</W> <W C='NNP'
L='w'>W</W> <W C='NN' L='p'>P</W> <W C='NNP' L='wong'>Wong</W> <W C='VBZ'
L='accept'>accepts</W> <W C='NNP' L='atv'>ATV</W> <W C='POS' L="'s">'s</W> <W
C='NN' L='apology'>apology</W> </S>
```

Figure 2: Sentence from document 19980303_004.e annotated with word boundaries and linguistic information.

those models to the NE task and outputs the same text with SGML marked-up text. The software was used with the pre-trained models and is available in both English and Chinese. The information about named entities in the corpus was kept in separate files and only used in some summarization experiments.

2.1.4. Sentence Alignment

Sentence alignment is the process of finding correspondences between source and target sentences in a pair of documents translation of each other. Sentence-level alignment was performed based on our re-implementation of Gate and Church's (1991) alignment algorithm. The basic assumption is that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. The information about sentence alignment is kept in tables and used in our cross-lingual evaluation.

2.2. Queries and Clusters

We used 400 documents for our experiments. They were clustered into document sets of 10 documents about one subject ("narcotics rehabilitation", "natural disaster victims aided", "customs staff doing good job", etc.). LDC annotators developed 40 such queries according to our guidelines, then they used an in-house information retrieval engine and human revision, to find the 10 most relevant documents for that query. We provided a manual Chinese translation of each query. Queries in English can be seen in Figure 4

2.3. Target Summaries

Three LDC judges then assessed each sentence in the 10 relevant documents, and assigned each sentence a score on a scale from 0 to 10, expressing how important this sentence is for the summary (Radev et al., 2000). This annotation, which is called "utility judgement", allows us to compile human-generated 'ideal' summaries at different compression rates, which is one gold-standard we use for our different measures of sentence-based agreement, both between the human agreement and between the system and the human annotators. We call this gold standard "human extracts". While utility judgement was only performed on English documents, in section 3.2. we show how we obtain human extracts for Chinese documents.

The judges also wrote multi-document summaries for each cluster at 50, 100, and 200 words (independently of the size of the documents). As human summary writing by trained professionals is very expensive, it was not possible to provide summaries of all 400 documents by several subjects (and several compression rates). However, our judges found the writing of multi-document summaries to be natural task. They followed the DUC guidelines to do so (DUC, 2000). These texts are a differ-

```
<?xml version="1.0"?>
<!DOCTYPE DOCSENT SYSTEM "/export/ws01summ/dtd/docsent.dtd" >
<DOCSENT DID="D-19980303_004.c" DOCNO="2203" LANG="CHIN" CORR-DOC="D-19980303_004.e">
<BODY>
<HEADLINE>
<S PAR="1" RSNT="1" SNO="1">  王永平 接納 亞洲 電視 道歉 </S>
</HEADLINE>
<TEXT>
<S PAR="2" RSNT="1" SNO="2">教育 統籌 局 局長 王永平 今日 （星期二） 表示，他 已 接納 亞洲
電視 有限 公司 對 其 在 上 星期 一（二月 二十二 日）「港 是 港 非」 節目 發表 的 評論 所 作
出 的 道歉，並 對 今次 事件，將 不 再 跟進。</S>
</TEXT>
</BODY>
</DOCSENT>
```

Figure 3: Document 19980303_004.c annotated with sentence boundaries.

TRAINING	
Group 125	Narcotics Rehabilitation
Group 241	Fire safety, building management concerns
Group 323	Battle against disc piracy
Group 551	Natural disaster victims aided
Group 112	Autumn and sports carnivals
Group 199	Intellectual Property Rights
Group 398	Flu results in Health Controls
Group 883	Public health concerns cause food-business closings
Group 1014	Traffi c Safety Enforcement
Group 1197	Museums: exhibits/hours
TEST	
Group 447	Housing (Amendment) Bill Brings Assorted Improvements
Group 827	Health education for youngsters
Group 885	Customs combats contraband/dutiable cigarette operations
Group 2	Meetings with foreign leaders
Group 46	Improving Employment Opportunities
Group 54	Illegal immigrants
Group 60	Customs staff doing good job.
Group 61	Permits for charitable fund raising
Group 62	Y2K readiness
Group 1018	Flower shows

Figure 4: 20 queries produced by the LDC.

ent gold standard we use (only for multi-document summaries); we call them "human summaries". Only human summaries in English are available in the corpus.

2.4. Inter-judge Agreement in Human Extracts

In order to measure agreement amongst the human extracts, we use Kappa (Siegel and Castellan, 1988), a statistical measure which addresses the problem of random agreement, and which is increasingly used in empirical NLP work and evaluation (Carletta, 1996). Kappa has the following advantages:

- It factors out random agreement. Random agreement is defined as the level of agreement which would be reached by random annotation using the same distribution of categories as the real annotators.
- It allows for comparisons between arbitrary numbers of annotators and items.
- It treats less frequent categories as more important (in our case: selected sentences), similarly to precision and recall but it also considers (with a smaller weight) more frequent categories as well.

The Kappa coefficient controls agreement $P(A)$ by taking into account agreement by chance $P(E)$:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

No matter how many items or annotators, or how the categories are distributed, $K = 0$ when there is no agreement other than what would be expected by chance, and $K = 1$ when agreement is perfect. If two annotators agree *less* than expected by chance, Kappa can also be negative. In our experiments we found low agreement among the judges (0.127 when selecting 10% of the text and 0.324 when selecting 90% of the text). This result is disappointing, but it is consistent with prior research (Rath et al., 1961) and it is still possible to use target summaries in different scenarios proposed in the literature (Salton et al., 1996; Mani, 2001).

2.5. Summarization Technologies and Automatic Summaries

During this workshop we have focused only on extractive summarization technology: automatic summaries are produced by selecting "relevant" sentences from the text representation. We have explored different summarization technologies that work on single and multi document mode. We have included two baseline methods in our framework: random summaries (constructed from sentences picked at random from the source) and lead based summaries (produced from sentences appearing on the beginning of the text). Random summaries should give a lower bound for the performance any system should have, while lead based summaries give a nice and simple baseline that sometimes obtain very good performance for specific tasks (see (Brandow et al., 1995)). More intelligent summarizers used in our evaluation are: Mead (Radev et al., 2000), Lexical Chains (Barzilay and Elhadad, 1997; Silber and McCoy, 2000), Summarist (Hovy and Lin, 1999), and Websumm (Mani and Bloedorn, 1999).

These summarization technologies were used to produce summaries for the entire corpus at different compression rates, covering many aspects of the summarization space (single, multi, cross-language summarization, word and sentence compression, query-based summarization) that were used for the purpose of evaluation of our experimental framework.

2.6. Information Retrieval Environment

We made use of an IR engine for conducting full-length document and summary retrieval. We adopted SMART (Salton, 1971) as the search engine for our retrieval experiments. The original SMART could only handle English documents. We changed the way it reads tokens so that it could deal with double-byte Chinese characters. We further configured the enhanced version of SMART, XSMART, to process the XML-formatted documents in the corpus.

For monolingual retrieval, queries are expressed in the same language as the documents. English and Chinese queries are used to retrieve the English and Chinese documents respectively. To obtain Chinese queries for conducting Chinese mono-lingual retrieval, the English queries were manually translated into Chinese queries by several native Chinese speakers on the team.

In addition to monolingual retrieval, we also explored cross-lingual retrieval for the relevance correlation measure that will be explained in the next section. In the cross-lingual retrieval setting, English queries are used for retrieving Chinese documents. Automatic query translation is applied to English queries producing Chinese queries. The automatically translated Chinese queries are then submitted to XSMART to retrieve Chinese documents.

3. Metrics for Evaluation in a Cross-lingual Environment

The evaluation of text summarization systems is an emergent research topic. Content evaluation assesses if automatic systems are able to identify the intended "topics" of the source document. Text quality evaluation assesses the readability, grammar and coherence of automatic summaries. Evaluations can be done in intrinsic or extrinsic fashions as defined by Sparck Jones and Galliers (1995). As part of our work we have explored both intrinsic and extrinsic evaluations. Our extrinsic evaluation assess summarization in a information retrieval task, our extrinsic evaluation assess the content of the summary by means of content-based similarity measures. In this paper we only focus on the description of the metrics for evaluation. The results of our experiments will be reported in detail elsewhere.

3.1. Information Retrieval Evaluation

One of the evaluation methods is to assess how well a summary supports information retrieval. We make use of a text retrieval engine to conduct retrieval on the English full-length documents for each query provided by LDC. The retrieval engine returns a ranked list of documents based on the relevance computed by the retrieval engine. Similarly, we conduct retrieval on the summaries instead of the full-length documents. Another ranked list of summaries are produced. The two ranked lists can then be compared. Relevance correlation (RC) is a new

measure for assessing the relative decrease in retrieval performance when moving from full documents to summaries.

There exist several methods for measuring the similarity of rankings. One such method is Kendall's tau and another is Spearman's rank correlation. Both methods are quite appropriate for the task that we want to perform; however, since search engines produce relevance scores in addition to rankings, we can use a stronger similarity test, linear correlation. When two identical rankings are compared, their correlation is 1. Two completely independent rankings result in a score of 0 while two rankings that are reverse versions of one another have a score of -1.

Relevance correlation r is defined as the linear correlation of the relevance scores (x and y) assigned by two different IR algorithms on the same set of documents or by the same IR algorithm on different data sets. Relevance scores are obtained using each of the 20 queries described in Figure 4.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

The monolingual retrieval can also be performed on Chinese documents, using the manually translated queries.

Besides mono-language retrieval, we also conduct cross-language retrieval. Given an English query, we first perform term translation by using an automated query translation technique to produce a translated Chinese query. Then, Chinese text retrieval is performed to retrieve a ranked list of Chinese documents. Given the same translated Chinese query, we also conduct retrieval on Chinese summaries and the system returns a ranked list of summaries. Different aspects of rank comparisons can be made, for instance, comparing the rankings of Chinese summaries produced by cross-language retrieval with English summaries produced by mono-language retrieval.

3.2. Content-based Evaluation

If intrinsic evaluation is performed by comparing extracted sentences to a set of "correct" extracted sentences, then co-selection is measured by precision, recall and F-score (Firmin and Chrzanowski, 1999). But these measures only consider sentence identity and not sentence content to carry out the comparison, which has the following negative effect: if two extracts consist of different sentences, whereby the sentences convey the same meaning, they are judged as very different by this measure, even though intuitively they would be judged as equivalent. As consequence of the fact that these measures consider only binary decisions (a sentence either is or is not in the extract), they ignore partially correct answers. Also, many researchers have opposed these measures; the generally accepted opinion is that there is no such thing as one ideal summary. Instead, a summary consists of a set of main ideas that should be conveyed (Jing et al., 1998; Jones and Paice, 1992)

The most extensive extrinsic evaluation of summarization systems was the TIPSTER SUMMAC evaluation (Mani et al., 1998). SUMMAC was extremely labour-intensive because of the need for assessors who had to read each of the full documents or extracts, which is a clear disadvantage of extrinsic measures of evaluation.

In our research we investigated measures for content evaluation based on the notion of vocabulary overlap. They are developed to palliate the problems with precision and recall. As they are completely automatic, they overcome the problems of task-based evaluations. These metrics are believed to be quite effective in determining the informativeness of a summary (Mani, 2001), and can be used in both extractive and non-extractive summarization, single and multi-document summarization. Recent research has shown how content-based evaluation can be carried out in automatic or semi-automatic fashion (Donaway et al., 2000; Paice and Oakes, 1999).

Content-based similarity measures are functions that take as arguments two text representations and compute a real value in the interval $[0..1]$, the value 1 means that the two texts are closely related while the value 0 means that the two texts are quite different. We have specified and implemented the following measures:

Cosine similarity is computed using the following formula (Salton, 1988):

$$cos(X,Y) = \frac{\sum x_i * y_i}{\sqrt{\sum (x_i)^2} * \sqrt{\sum (y_i)^2}}$$

where X and Y are text representations based on a vector space model. We use two possible weighting schemes for the terms: presence/absence of the term in the text or $tf * idf$ computed using corpus and within text term distribution.

Unit overlap is computed using the following formula:

$$overlap(X,Y) = \frac{\|X \cap Y\|}{\|X\| + \|Y\| - \|X \cap Y\|}$$

where X and Y are text representations based on sets. Here $\|S\|$ is the size of set S.

Longest Common Subsequence is computed using the formula:

$$2 * lcs(X,Y) = length(X) + length(Y) - edit_{di}(X,Y)$$

where X and Y are representations based on sequences and where $lcs(X,Y)$ is the length of the longest common subsequence between X and Y, $length(X)$ is the length of the string X, and $edit_{di}(X,Y)$ is the minimum number of deletion and insertions needed to transform X into Y (Crochemore and Rytter, 1994). When comparing two texts, we compute a normalized pairwise lcs between the sentences of the two texts. Unlike cosine and overlap, longest common subsequence is sensitive on how information is sequenced in the text. Content-based similarity measures have been used in the past to assess machine translation quality (Papineni et al., 2001).

Different measures require different text representations: cosine is based on the vector space model, while unit overlap is based on a set data type and longest common subsequence operates in the sequence data type. One can compare text units at different levels of analysis: For example one can compare units relying on the number of word or token that two units share, or one can compare the number of lemmas they share. One can use only nouns as the representation, based on the idea that are the nouns that carry the content of the sentence; one might alternatively use main verbs. We experimented with all these parameters and allow our measures to operate at different granularity levels. For each automatic extract, one can compute its average similarity to a set of target extracts. Further, for each extract one can compute its maximum and minimum similarity to a set of target extracts.

The experimental framework for evaluation of the Chinese summaries is based on the novel idea of using the aligned corpus as a source for obtaining a target abstract in Chinese. Given a collection of monolingual summaries, we can use our alignment tables to generate reasonable corresponding cross-lingual summaries and use the collection of these "pseudo manual" Chinese summaries in our experiments. This was at all possible because of the accuracy of the alignment program: A preliminary evaluation of our alignment algorithm measured precision and recall at 95.5% each.

We have based this evaluation on human extracts produced by LDC assessors (and sentence-alignment in the Chinese case). Nevertheless, other alternatives exist: Content-based similarity measures do not require the target summary to be a subset of sentences from the source document, thus, content evaluation based on similarity measures can be done using human-written summaries. In our experiments, we have compared human multi-document extracts with human multi-document summaries. We have also compared automatic multi-document summaries with human multi-document summaries. Our experiments show the use of our framework for comparing human and automatic extracts with human *abstracts*, i.e. coherent, newly written summaries of the documents rather than sentence extracts.

4. Conclusions

In this paper we have described the development of language and processing resources for the evaluation of automatic summarization systems. From the point of view of the data, we have sentence-aligned and annotated a pre-existing parallel corpus of English and Chinese documents, developed queries in both languages, and manually constructed clusters of documents relevants to each query. We have also provided with sentence relevance measures for each sentence in the document clusters, constructed automatic extracts using different methods, and constructed manual multi-document summaries. From the point of view of the software components, we have developed tools for the evaluation of text summarization systems and provided with baseline and one modular state-of-the art summarizer, Mead, that produces single-document, multi-document, generic, and query-based summaries. Our work provides data and tools for evaluation of extractive, non-extractive, single and multi-document summarization. All resources are being made available to the research community (http://www.clsp.jhu.edu/ws2001/groups/asmd).

Acknowledgements

We are grateful to Arda Celebi, John Blitzer, Hong Qi, Danyu Liu, and Elliot Drabek for their work during the workshop. We thank Fred Jelinek, Sanjeev Khudanpur and the staff of the Center for Language and Speech Processing, Johns Hopkins University for their hospitality. We are grateful to Inderjeet Mani. We also thank Chin-Yew Lin, Greg Silber, and Regina Barzilay. We would like to thank Hamish Cunningham for providing the English morphological analyser. The 2001 Summer Workshop at Johns Hopkins University was sponsored by the National Science Foundation via Grant No. IIS-0097467, which included support from the Defense Advanced Research Projects Agency.

5. References

Regina Barzilay and Michael Elhadad. 1997. Using Lexical Chains for Text Summarization. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent*

Scalable Text Summarization, pages 10–17, Madrid, Spain, July.

GTE. BBN Technologies, 2000. *IdentiFinder: User Manual*, July. Version 5.0.

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

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *CL*, 22(2):249–254.

M. Crochemore and W. Rytter. 1994. *Text Algorithms*. Oxford University Press.

R.L. Donaway, K.W. Drummey, and L.A. Mather. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *Proceedings of the Workshop on Automatic Summarization, ANLP-NAACL2000*, pages 69–78. Association for Computational Linguistics, 30 April 2000.

2000. *Document Understanding Conference*.

T. Firmin and M.J. Chrzanowski. 1999. An Evaluation of Automatic Text Summarization Systems. In I. Mani and M.T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 325–336. The MIT Press.

W.A. Gale and K.W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpus. In *COLIN 91*, pages 177–184.

C. Grover, C. Matheson, A. Mikheev, and M. Moens. 2000. LT TTT: A Flexible Tokenisation Tool. In *Proceedings of LREC'00*.

E. Hovy and C-Y. Lin. 1999. Automated Text Summarization in SUMMARIST. In I. Mani and M.T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 81–94. The MIT Press.

K. Humphreys, R. Gaizauskas, and H. Cunningham. 2000. LaSIE Technical Specifications. Technical report, Department of Computer Science. University of Sheffield.

Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *AAAI 98 Spring Symposium on Intelligent Text Summarization*, pages 60–68.

P.A. Jones and C.D. Paice. 1992. A 'select and generate' approach to to automatic abstracting. In A.M. McEnry and C.D. Paice, editors, *Proceedings of the 14th British Computer Society Information Retrieval Colloquium*, pages 151–154. Springer Verlag.

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

I. Mani, D. House, G. Klein, L. Hirshman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. 1998. The TIPSTER SUMMAC Text Summarization Evaluation. Technical Report Technical Report MTR 98W0000138, The Mitre Corporation, McLean, Virginia.

Inderjeet Mani. 2001. *Automatic Summarization*. John Benjamins Publishing Company.

A. Mikheev. 2000. Tagging Sentence Boundaries. In *Proceedings of the NAACL*, Seattle, USA. ACL.

G. Miller. 1995. WordNet: a Lexical Database for English. *Communications of the ACM*, Volume 38(Number 11), November.

C.D. Paice and M.P. Oakes. 1999. A Concept-Based Method for Automatic Abstracting. Technical Report Research Report 27, Library and Information Commission.

K. Papineni, S. Rouskos, T. Ward, and W-J. Zhu. 2001. Blue: A Method for Automatic Evaluation of Machine Translation. Research Report RC22176, IBM.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, WA, April.

G. Rath, A. Resnick, and R. Savage. 1961. The formation of abstracts by the selection of sentences: Part 1: sentence selection by man and machines. *American Documentation*, 12(2):139–141.

Gerard Salton, James Allan, and Amit Singhal. 1996. Automatic text decomposition and structuring. *Information Processing & Management*, 32(2):127–138.

G. Salton. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice Hall, Englewood Cliffs, NJ.

G. Salton. 1988. *Automatic Text Processing*. Addison-Wesley Publishing Company.

Sidney Siegel and N. John Jr. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Berkeley, CA, 2nd edition.

H.G. Silber and K.F. McCoy. 2000. Efficient text summarization using lexical chains. In *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI'2000)*.

K. Sparck Jones and J.R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review*. Number 1083 in Lecture Notes in Artificial Intelligence. Springer.