Generic Sentence Fusion is an Ill-Defined Summarization Task

Hal Daumé III and **Daniel Marcu**
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
{hdaume,marcu}@isi.edu

Abstract

We report on a series of human evaluations of the task of sentence fusion. In this task, a human is given two sentences and asked to produce a single coherent sentence that contains only the *important* information from the original two. Thus, this is a highly constrained summarization task. Our investigations show that even at this restricted level, there is no measurable agreement between humans regarding what information should be considered important. We further investigate the ability of separate evaluators to assess summaries, and find similarly disturbing lack of agreement.

1 Introduction and Motivation

The practices of automatic summarization vary widely across many dimensions, including source length, summary length, style, source, topic, language, and structure. Most typical are summaries of a single news document down to a headline or short summary, or of a collection of news documents down to a headline or short summary (Hahn and Harman, 2002). A few researchers have focused on other aspects of summarization, including single sentence (Knight and Marcu, 2002), paragraph or short document (Daumé III and Marcu, 2002), query-focused (Berger and Mittal, 2000), or speech (Hori et al., 2003).

The techniques relevant to, and the challenges faced in each of these tasks can be quite different. Nevertheless, they all rely on one critical assumption: there exists a notion of (relative) importance between pieces of information in a document (or utterance), regardless of whether we can detect this or not. Indeed, recent research has looked at this question in detail, and can be rather cleanly divided into two partitions. The first partition aims to develop *manual* evaluation criteria for determining the quality of a summary, and is typified by the extensive research done in single-document summarization by Halteren and Teufel (2003) and by the evaluation strategy proposed by Nenkova and Passonneau (2004). The other half aims to develop *automatic* evaluation criteria to imitate the manual evaluation methods (or at least to complement them). Work in this area includes that of Lin and Hovy (2003) and Pastra and Saggion (2003), both of whom inspect the use of Bleu-like metrics (Papineni et al., 2002) in summarization.

The results of these investigations have been mixed. In the DUC competitions (Hahn and Harman, 2002), when manual evaluation has been employed, it has been commonly observed that human-written summaries grossly outscore any machine-produced summary. All machine-produced summaries tend to show little (statistically significant) difference from one another. Moreover, a baseline system that simply takes the first sentences of a document performs just as well or better than intelligently crafted systems when summarizing news stories. Additionally, studies of vast numbers of summaries of the same document (Halteren and Teufel, 2003) have shown that there is little agreement among different humans as to what information belongs in a single document summary. This has been leveraged by Nenkova and Passonneau (2004) to produce a manual scoring method for summaries, though the fact that humans show so little agreement in this task is somewhat disheartening. All of these evaluations rely strongly on the issue of multiple references, in order to achieve consensus.

Opinions voiced at DUC meetings indicate that different researchers attribute this apparent lack of agreement to one (or more) of many factors (in addition, see (Mani and Maybury, 1999)). Many believe that the fact that we are typically working in a news genre is to blame, though this complaint tends to be directed more at the excellent performance of the baseline than at the issue of human agreement. Others

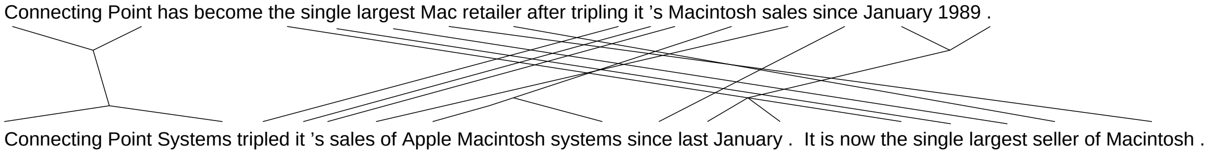
Connecting Point has become the single largest Mac retailer after tripling it 's Macintosh sales since January 1989 .
Connecting Point Systems tripled it 's sales of Apple Macintosh systems since last January . It is now the single largest seller of Macintosh .

Figure 1: Example ⟨document, abstract⟩ alignment.

believe that in order to observe more agreement, one needs to move to query-focused summaries; it seems reasonable that if the person writing the summary knew how it would be used, he would be more guided in what information to retain. Yet others attribute the lack of agreement simply to the vast space of possible choices a summarizer could make, and see the disagreement simply as par for the course.

2 Our Study

In this paper, we report on a study of the performance of humans producing summaries. We concern ourselves with the task of *sentence fusion.* In this task, we assume that two sentences are provided and that the summarizer must produce as output a single sentence that contains the important information contained in the input sentences (we will describe later how we obtain such data). We would like to show that this task is well-defined: if we show many humans the same two sentences, they will produce similar summaries. Of course we do not penalize one human for using different words than another.

The sentence fusion task is interesting after performing sentence extraction, the extracted sentences often contain superfluous information. It has been further observed that simply compressing sentences individually and concatenating the results leads to suboptimal summaries (Daumé III and Marcu, 2002). The use of sentence fusion in *multi-document* summarization has been extensively explored by Barzilay in her thesis (Barzilay, 2003; Barzilay et al., 1999), though in the multi-document setting, one has redundancy to fall back on. Additionally, the sentence fusion task is sufficiently constrained that it makes possible more complex and linguistically motivated manipulations than are reasonable for full document or multi-document summaries (and for which simple extraction techniques are unlikely to suffice).

3 Data Collection

Our data comes from a collection of computer product reviews from the Ziff-Davis corporation. This corpus consists of roughly seven thousand documents paired with human written abstracts. The average document was 1080 words in length, with an abstract of length 136 words, a compression rate of roughly 87.5%.

3.1 Examples Based on Alignments

For 50 of these ⟨document, abstract⟩ pairs, we have human-created word-for-word and phrase-for-phrase alignments. An example alignment is shown in Figure 1. Moreover, using a generalization of a hidden Markov model, we are able to create (in an unsupervised fashion) similar alignments for all of the documents (Daumé III and Marcu, 2004). This system achieves a precision, recall and f-score of 0.528, 0.668 and 0.590, respectively (which is a significant increase in performance (f = 0.407) over the IBM models or the Cut & Paste method (Jing, 2002)).

Based on these alignments (be they manually created or automatically created), we are able to look for examples of sentence fusions within the data. In particular, we search for sentences in the abstracts which are aligned to exactly two document sentences, for which at least 80% of the summary sentence is aligned and for which at least 20% of the words in the summary sentence come from each of the two document sentences.

This leaves us with pairs that consist of two document sentences and one abstract sentence, exactly the sort of data we are looking to use. We randomly select 25 such pairs from the data collected from the human-aligned portion of the corpus and 25 pairs from the automatically aligned portion, giving us 50 pairs in all.

3.2 Examples Based on Elicitation

In addition to collecting data from the Ziff-Davis corpus, we also elicited data from human subjects with a variety of different backgrounds (though all were familiar with computers and technology). These people were presented with the pairs of document sentences and, *independently of the rest of the document,* asked to produce a single summary sentence that contained the "important" information. Their summary was to be about half the length of the original

ORIG:	After years of pursuing separate and conflicting paths, AT&T and Digital Equipment Corp. agreed in June to settle their computer-to-PBX differences. The two will jointly develop an applications interface that can be shared by computers and PBXs of any stripe.
REF:	AT&T and DEC have a joint agreement from June to develop an applications interface to be shared by various models of computers and PBXs.
HUM 1:	AT&T and Digital Equipment Corp. agreed in June to settle their computer-to-PBX differences and develop an applications interface that can be shared by any computer or PBX.
HUM 2:	After years of pursuing different paths, AT&T and Digital agreed to jointly develop an applications interface that can be shared by computers and PBXs of any stripe.
HUM 3:	After working separately for years, AT&T will jointly develop an interface between computers and PBXs.

Table 1: Example of elicited data.

(this is what was observed in the pairs extracted from the corpus) They were given no additional specific instructions.

The summaries thus elicited ranged rather dramatically from highly cut and paste summaries to highly abstractive summaries. An example is shown in Table 1. In this table, we show the original pair of document sentences, the "reference" summary (i.e., the one that came from the original abstract), and the responses of three of the eight human subjects are shown (the first is the most "cut and paste," the second is typical of the "middle set" and the last is unusually abstractive).

3.3 Baseline Summaries

In addition to the human elicited data, we generate three baseline summaries. The first baseline, LONGER, simply selects the longer of the two sentences as the summary (typically the sentences are roughly the same length; thus this is nearly random). The second baseline, DROPSTOP first catenates the sentences (in random order), then removes punctuation and stop words, finally cutting off at the 50% mark. The third baseline, COMP is the document compression system developed by Daumé III and Marcu (2002), which compresses documents by cutting out constituents in a combined syntax and discourse tree.

4 Evaluation of Summaries

We perform three types of manual evaluation on the summaries from the previous section. In the first, the **ranked evaluation**, we present evaluators with *original* two document sentences; they also see a list of hypothesis summaries and are asked to rank them relative to one another. In the second evaluation, the **absolute evaluation**, evaluators are presented with the reference summary and a hypothesis and are asked to produce an absolute score for the hypothesis. In the third, the **factoid evaluation**, we manually inspect the information content of each hypothesis.

4.1 Ranked Evaluation

In the ranked evaluation, human evaluators are presented with the original two document sentences. They also see a list of 12 hypothesis summaries: the reference summary, the eight summaries elicited from human subjects, and the three baseline summaries. They are asked to produce a ranking of the 12 summaries based both on their faithfulness to the original document sentences and on their grammaticality. They were allowed to assign the same score to two systems if they felt neither was any better (or worse) than the other. They ranked the systems from 1 (best) to 12 (worst), though typically enough systems performed "equally well" that a rank of 12 was not assigned. Three humans performed this evaluation.

4.2 Absolute Evaluation

In the absolute evaluation, human evaluators are shown the reference summary and a single hypothesis summary. In order to partially assuage the issue of humans doing little more than string matching (Coughlin, 2001), the reference and hypothesis were shown on separate pages and humans were asked not to go "back" during the evaluation. Due to time constraints, only three systems were evaluated in this manner, one of the humans (the human output was selected so that it was neither too cut-and-paste nor too generative), the LONGER and COMP systems. Three humans performed this task (each shown a single different system output for each reference summary) and scored outputs on a scale from 1 (best) to 5 (worst). They were told to deduct points for any information con-

Ref	Longer	Comp	Hum 1	Hum 2	Hum 3	Factoid
	★	★	★	★	★	CP has taken leadership
★					★	leadership by volume
						doug kass is analysis at dataquest inc
						dq is a market research co
						dq is in san jose
						kass said CP has taken leadership
★	★	★	★	★		analysts say
★	★	★	★	★	★	CP has a wide variety of stores
★	★	★	★	★	★	CP endorsed apple's earned investment program
★	★	★	★			CP has become the low-price leader
★	★	★	★			CP hasn't sacrificed technical support

Table 2: Factoid-based evaluation scheme for the sentence pair "Connecting Point has taken leadership by volume, volume, volume," said Doug Kass, an analyst at Dataquest Inc., a market research company in San Jose. Analysts and observers say Connecting Point's wide variety of stores and endorsement of Apple's earned investment program have helped it become the low-price leader without sacrificing technical support."

tained in the reference not contained in the hypothesis, any information contained in the hypothesis not contained in the reference, and ungrammaticality.

4.3 Factoid Evaluation

The third evaluation we perform ourselves, due to its difficulty. This follows the general rubric described by Nenkova and Passonneau's (2004) pyramid scoring scheme, though it differs in the sense that we base our evaluation not on a reference summary, but on the original two document sentences. Our methodology is described below.

We assume that we are given the original pair of sentences from the document and the hypothesis summaries for many systems (in our experiments, we used the original reference summary, the outputs of three representative humans, and the Longer and Comp baselines). Given this data, we first segment the original pair of sentences into "factoids" in the style of Halteren and Teufel (2003). Then, for each hypothesis summary and each factoid, we indicate whether the summary contained that factoid.

Grammaticality of summary hypotheses enters into the calculation of the factoid agreement numbers. A system only gets credit for a factoid if its summary contains that factoid in a sufficiently grammatical form that the following test could be passed: given any reasonable question one could pose about this factoid, and given the hypothesis summary, could one answer the question correctly. An example is shown in Table 2.

Based on this information, it is possible to select one or more of the outputs as the "gold standard" and compare the rest in the pyramid scoring scheme described by Nenkova and Passonneau (2004). If only one output is used as the gold standard, then it is sufficient to compute precision and recall against that gold standard, and then use these numbers to compute an F-score, which essentially measures agreement between the chosen gold standard and another hypothesis. In the remainder of this analysis, when we report an F-score over the factoid, this is calculated when the Ref summary is taken as the standard.

	Hum 1	Hum 2	Hum 3
Ref	0.182	0.188	0.251
Hum 1	-	0.201	0.347
Hum 2	-	-	0.470

Table 3: Agreement (kappa) scores for different combinations of systems and humans.

5 Evaluation Results

The fundamental question we would like to answer is whether humans agree in terms of what information should be preserved in a summary. Given our data, there are two ways of looking at this. First: do the humans from whom we elicited data select the same information as the reference? Second: do these humans agree with each other. Both of these questions can be answered by looking at the results of the factoid evaluation.

For any set of columns in the factoid evaluation, we can compute the agreement based on the kappa statistic (Krippendorff, 1980). Researchers have observed that kappa scores over 0.8 indicate strong agreement, while scores between 0.6 and 0.8 indicate reasonable agreement. Kappa values below 0.6 indicate little

System	F-Score	Absolute	Relative
Hum 4	0.652	2.605	2.066
Hum 3	0.608	-	2.276
Hum 5	0.574	-	2.434
Longer	0.419	3.000	3.368
Ref	1.000	-	3.500
Comp	0.475	3.842	4.184

Table 4: Factoid F-score, absolute score and relative ranking for 6 outputs.

to no agreement. The kappa values for various combinations of columns are shown in Table 3.

As we can see from this table, there is essentially no agreement found anywhere. The maximum agreement is between Human 2 and Human 3, but even a kappa value of 0.470 is regarded as virtually no agreement. Furthermore, the kappa values comparing the human outputs to the reference outputs is even lower, attaining a maximum of 0.251; again, no agreement. One is forced to conclude that in the task of generic sentence fusion, people will not produce a summary containing the same information as the original reference sentence, and will not produce summaries that contain the same information as another person in the same situation.

Despite the fact that humans do not agree on what information should go into a summary, there is still the chance that when presented with two summaries, they will be able to distinguish one as somehow better than another. Answering this question is the aim of the other two evaluations.

First, we consider the absolute rankings. Recall that in this evaluation, humans are presented with the reference summary as the gold standard summary. Since, in addition to grammaticality, this is supposed to measure the correctness of information preservation, it is reasonable to compare these numbers to the F-scores that can be computed based on the factoid evaluation. These results are shown in Table 4. For the first column (F-Score), higher numbers are better; for the second and third columns, lower scores are better. We can see that the evaluation prefers the human output to the outputs of either of the systems. However, the factoid scoring prefers the Comp model to the Longer model, though the Absolute scoring rates them in the opposite direction.

As we can see from the Relative column in Table 4, human elicited summaries are consistently preferred to any of the others. This is good news: even if people cannot agree on what information should go into a summary, they at least prefer human written summaries to others. After the human elicited summaries, there is a relatively large jump to the Longer baseline, which is unfortunately preferred to the Reference summary. After the reference summary, there are two large jumps, first to the document compression model and then to the DropStop baseline. However, when comparing the relative scores to the F-Score, we see that, again, the factoid metric prefers the Comp model to the Longer model, but this is not reflected in the relative scoring metric.

6 Analysis of Results

There are two conclusions that can be drawn from these data. The first, related specifically to the kappa statistic over the factoids as depicted in Table 3, is that even in this modest task of compressing two sentences into one, the task is ill-defined. The second, related to the two other evaluations, is that while humans seem able to agree on the relative quality of sentence fusions, judgments elicited by direct comparison do not reflect whether systems are correctly able to select content.

6.1 Disagreement of Importance

As indicated in Section 5, when humans are given the task of compressing two sentences into one, there is no measurable agreement between any two as to what information should be retained.

The first thing worth noting is that there is moderately more agreement between two elicited, non-expert data points than between the elicited data and the original reference. This can be attributed either to the lack of context available to the non-experts, or to their respective lack of expertise. Regardless, the level of agreement between such non-expert humans is so low that this matters little. Furthermore, from an automatic sentence fusion perspective, a computer program is much more like a non-expert human with no context than an expert with an entire document to borrow from.

It might be argued that looking at only two sentences does not provide sufficient context for humans to be able to judge relative importance. This argument is supported by the fact that, upon moving to multi-document summarization, there is (relatively) more agreement between humans regarding what pieces of information should be kept. In order to make the transition from two-sentence fusion to multi-document summarization, one essentially needs to make two inductive steps: the first from two sentences, to three and so on up to a full *single* document; the second from a single document to multiple documents.

The analysis we have performed does not comment on either of these inductive steps. However, it is much more likely that it is the second, not the first, that breaks down and enables humans to agree more when creating summaries of collections of documents. On the one hand, it seems unreasonable to posit that there is some "magic" number of sentences needed, such that once two humans read that many sentences, they are able to agree on what information is relevant. On the other hand, in all evaluations that have considered multi-document summarization, the collection of documents to be summarized has been selected by a human with a particular interest in mind. While this interest is not (necessarily) communicated to the summarizers directly, it is indirectly suggested by the selection of documents. This is why the use of redundancy in multi-document summarization is so important. If, on the other hand, humans were given a set of moderately related or unrelated documents, we believe that there would be even less agreement on what makes a good summary[1].

6.2 Human Perception of Quality

We have presented two sets of results regarding human perception of the quality of summaries. In the first (see Table 4), humans are presented with the Ref summary and then with either a human-elicited summary, a summary that is simply the longer of the two sentences (recall that they *do not* see the original two sentences, so they have no way of knowing how this summary was created) and the output of the Comp system. If one accepts that the F-Score over factoids is a high-quality measure of summary quality, then there should be strong correlation between this F-Score and the absolute scoring of the system outputs. This is not observed. In fact, the F-Score strongly prefers the Comp system over the Longer system, while human scoring prefers the Longer system.

Since the humans performing this evaluation were told explicitly to count off for missing information, extraneous information or lack of grammaticality, the only reasonable explanation for this discrepancy is that the evaluators were sufficiently put off by the grammatical errors made by the Comp system that they penalized it heavily. Grammaticality does enter into the factoids evaluation, though perhaps not as strongly.

In the relative ranking evaluation (see Table 4), there are two disturbing observations we can make. First, as in the absolute scoring, the factoid evaluation prefers the Comp system to the Longer system, but the relative ranking puts them in the other order. Second, the Longer baseline *outperforms* the reference summary.

As before, we can explain this first discrepancy by the issue of grammaticality. This is especially important in this case: since the evaluators are not given a reference summary that explicitly tells them what information is important and what information is not, they are required to make this decision on their own. As we have observed, this act is very imprecise, and it is likely the people performing the evaluation have recognized this. Since there is no longer a clear cut distinction between important and unimportant information, and since they are required to make a decision, they have no choice but to fall back on grammaticality as the primary motivating factor for their decisions.

The second discrepancy is particularly disturbing. Before discussing its possible causes, we briefly consider the implications of this finding. In order to build an automatic sentence fusion system, one would like to be able to automatically collect training data. Our method for doing so is by constructing word-for-word and phrase-for-phrase alignments between documents and abstracts and leveraging these alignments to select such pairs. In theory, one could extract many thousands of such examples from the plethora of existing document/summary pairs available. Unfortunately, this result tells us that even if we are able to build a system that perfectly mimics these collected data, a simple baseline will be preferred by humans in an evaluation.

One might wish to attribute this discrepancy to errors made by the largely imperfect automatic alignments. However, we have calculated the results separately for pairs derived from

[1] Summarizing a set of unrelated documents may be an unrealistic and unimportant task; nevertheless, it is interesting to consider such a task in order to better understand why humans agree more readily in multi-document summarization than in single document summarization or in sentence fusion.

human alignments and from automatic alignments, and observe no differences.

This leaves two remaining factors to explain this difference. First, the original summary is created by a trained human professional, who is very familiar with the domain (while our elicited data comes from technologically proficient adults, the topics discussed in the data are typically about technical systems from the late eighties, topics our summarizers know very little about). Second, the original summarizers had the rest of the document available when creating these fusions. Though without performing relevant experiments, it is impossible to say what the results would be.

However, from a system-building perspective, one can view fusion in many applications and it is highly desirable to be able to perform such fusions without knowing the rest of the document. From a document summarization perspective, one might wish to perform sentence extraction to reduce the document to a few sentences and then use sentence fusion to compress these further. In this case, the primary motivation for performing this in a pipelined fashion would be to remove the complexity of dealing with the entire document when the more complex fusion models are applied. In another possible application of question answering, one can imagine answering a question by fusion together several sentences returned as the result of an information retrieval engine. In this case, it is nearly impossible to include the remainder of the documents in such an analysis.

7 Summary and Conclusions

We have performed an analysis of agreement between humans in the highly constrained task of fusing two sentences together. This task has applications in summarization, question answering and pure natural language generation. We have shown that this task is *not* well defined, when viewed in isolation. Furthermore, we have shown that using automatically extracted data for training cannot lead to systems that outperform a simple baseline of choosing the longer of the two sentences..

These results are disheartening, though by performing such experiments a priori, we are able to better judge which courses of research are and are not worth pursuing. Questions regarding the agreement between people in the area of single document summarization and multi-document summarization have already been raised and are currently only partially answered (Halteren and Teufel, 2003; Nenkova and Passonneau, 2004; Marcu and Gerber, 2001). We have shown that even in this constrained domain, it is very unlikely that any significant agreement will be found, without specifically guiding the summarizers, either by a query, a user model, or some other external knowledge. We have argued that it is likely that this lack of agreement will not be subverted by adding more sentences, though this should be confirmed experimentally.

The issues of multiple references and of adding context (essentially by allowing the summarizers to see the document from which these two sentences were extracted) has not been addressed in this work; either might serve to increase agreement. However, one of the goals of this methodology for automatically extracting pairs of sentences from automatically aligned corpora is to be able to get data on which to train and test a system without having humans write it. To require one to elicit multiple references to obtain any agreement obviates this goal (moreover, that agreement between humans and the original summary sentence is even lower than between a pair of humans makes this practice questionable). Regarding context, it is reasonable to hypothesize (though this would need to be verified) that the addition of context would result in higher kappa scores. Unfortunately, if a human is given access to this information, it would only be fair to give a system access to the same information. This means that we would no longer be able to view generic sentence fusion as an isolated task, making fusion-specific research advances very difficult.

Acknowledgements

We wish to thank Kevin Knight, Eduard Hovy, Jerry Hobbs and the anonymous reviewers for their helpful and insightful comments. This work was partially supported by DARPA-ITO grant N66001-00-1-9814, NSF grant IIS-0097846, and a USC Dean Fellowship to Hal Daumé III.

References

R. Barzilay, K. McKeown, and M. Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proc. of ACL.*

R. Barzilay. 2003. *Information Fusion for Mutlidocument Summarization: Paraphras-*

ing and Generation. Ph.D. thesis, Columbia University.

A. Berger and V. Mittal. 2000. Query-relevant summarization using FAQs. In *Proc. of ACL.*

D. Coughlin. 2001. Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit IX.*

H. Daumé III and D. Marcu. 2002. A noisy-channel model for document compression. In *Proc. of ACL.*

H. Daumé III and D. Marcu. 2004. A phrase-based HMM approach to document/abstract alignment. *In preparation.*

U. Hahn and D. Harman, editors. 2002. *Second Document Understanding Conference (DUC-2002).*

H. Halteren and S. Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *In HLT-NAACL DUC Workshop.*

C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel. 2003. A statistical approach to automatic speech summarization. *Journal on Applied Signal Processing*, 3:128–139.

H. Jing. 2002. Using hidden Markov modeling to decompose human-written summaries. *Computational Linguistics*, 28(4):527 – 544, December.

K. Knight and D. Marcu. 2002. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence.*

K. Krippendorff. 1980. *Content analysis: An Introduction to its Methodology.* Sage Publications, CA.

C.Y. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. of HLT-NAACL.*

C.Y. Lin. 2003. Improving summarization performance by sentence compression - a pilot study. In *Proc. of IRAL Workshop.*

I. Mani and M. Maybury, editors. 1999. *Advances in Automatic Text Summarization.* The MIT Press, Cambridge, MA.

D. Marcu and L. Gerber. 2001. An inquiry into the nature of multidocument abstracts, extracts, and their evaluation. In *NAACL Summarization Workshop.*

A. Nenkova and R. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL.*

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL.*

K. Pastra and H. Saggion. 2003. Colouring summaries BLEU. In *In EACL.*