Task-Based Evaluation of Summary Quality: Describing Relationships Between Scientific Papers

Simone Teufel
Computer Science Department
Columbia University
New York, NY-10027
teufel@cs.columbia.edu

Abstract

We present a novel method for task-based evaluation of summaries of scientific articles. The task we propose is a question-answering task, where the questions are about the relatedness of the current paper to prior research. This evaluation method is time-efficient with respect to material preparation and data collection, so that it is possible to test against many different baselines, something that is not usually feasible in evaluations by relevance decision. We use this methodology to evaluate the quality of summaries our system produces. These summaries are designed to describe the contribution of a scientific article in relation to other work. The results show that this type of summary is indeed more useful than the baselines (random sentences, keyword lists and generic author-written summaries), and nearly as useful as the full texts.

1 Introduction

Extrinsic or *task-based* evaluation of summary quality is considered by many as one of the best forms of evaluation: the value of a summary, as a functional text, lies in how well it serves to fulfill a function. In task-based evaluation, the quality of a summary is measured as the task performance it enables a user to achieve. In contrast, intrinsic evaluation measures properties of the summary in isolation: how concise, syntactically well-formed, coherent or information-preserving it is (Jing and McKeown, 2000) or to which degree a summary resembles an "ideal" summary or gold standard (i.e. an a-priori definition of what the summary should look like). Even though extrinsic evaluation requires much more effort than intrinsic ones, interest in extrinsic evaluation is growing in the summarisation community.

Typically, the task used in extrinsic summary evaluation is *relevance decision* in an information retrieval (IR) context. Given a query and a set of documents, the subjects have to decide for each document if it is relevant to the query. In one experimental condition, the documents are represented by their summaries only, in the other by the full texts. The two variables measured are task completion time and task performance (e.g. recall and precision of correct relevance decisions). Brevity being the main advantage of a summary, the perfect summary is one which allows a user to predict the relevance of a document to a query as well as the full paper would have, while saving reading time.

Sentence extracts, one simple form of summaries, provide enough information for subjects to perform informed relevance decisions: Tombros et al. (1998) found that their query-based sentence extracts improved recall on 50 TREC queries from 49.8% to 65.6% when compared to typical IR output (namely title and first few sentences) and precision from 44.3% to 55.3%. Their sentence extracts also increased speed: users were able to examine 22.6 documents in 5 minutes, compared with 20 documents. Mani et al. (1999a), evaluating 16 sentence-extraction-based systems contrastively in the large-scale TIPSTER SUMMAC evaluation exercise, found that summaries as short as 17% of the full text length can speed up decision making by a factor of 2, without degrading F-score accuracy.

However, there is evidence that something *simpler* than sentence extracts might also work well for relevance decision. For example, we know that experts in a field often decide on the basis of title and author information alone if they need to read a paper or not (Bazerman, 1985); this is particularly the case in medicine where the titles tend to be long and informative. Indexing of documents provides another informative document surrogate, whereby the semantics of a document is described by keywords, chosen by a human or automatically (eg. by the TF*IDF formula). Yet another example for a simple baseline are random sentences. However, previous task-based summary evaluations have not always compared performance against these kinds of simpler baselines, possibly due to the extensive effort required to prepare and run relevance decision evaluations.

Summaries are sophisticated texts: far from being just collections of keywords, they are coherent texts expressing connected facts. It requires considerably more effort to produce summaries than simpler representations of the text. This is particularly the case for a new generation of summaries, which are generated out of sentence parts, e.g. by shortening, fusing or other forms of revision (Grefenstette, 1998; Mani et al., 1999b; Barzilay et al., 1999; Jing and McKeown, 2000; Knight and Marcu, 2000). While simpler representations like keywords often accurately portray the topic of a text, the added value of summaries lies in their ability to convey more complex information about concepts and events and their relation to the overall message of the document.

We believe that relevance decisions do not fully measure this added value of summaries. It is much harder to decide *in which respect* two documents are related than to decide what the *topic* of a document is. Arguably, such a harder task is needed to show an advantage of summaries over simpler document surrogates. The practical problem is that it is not trivial to pinpoint new tasks well enough to use them in a formal evaluation.

We propose one such task here, i.e. the task of deciding in which respect a scientific paper relates to previous work it describes and cites. In section 2, we will explain in which way the task of determining relationships between papers is relevant to researchers.

The evaluation (cf. section 5) is performed as follows: Subjects are given different representations of a paper and then they are asked to determine which of the approaches mentioned in the paper are criticised and which approaches are used in a supportive fashion. Their performance on this task is measured in terms of number of correct answers. In the first instance, we are interested in how well humans can perform the task if given full information (i.e., the full text); this measures if the task is well-defined.

We then create conditions in which subjects have access to substantially less information; their task performance, compared to their performance with the full paper, is taken as a measure of the usefulness of the given document representation. In our evaluation methodology, it is possible to collect data points for many different baselines because data collection is very time-efficient. The document representations shown to subjects are either the author-written summary, or a list of keywords, or a list of randomly chosen sentences, or summaries created by our system (Teufel and Moens, 2000). The focus of this paper is the evaluation methodology, not the system; however a brief overview of the system is given in section 4.

The results of our evaluation (cf. section 5.2) show that simple baselines, including the author-supplied summaries, do *not* provide sufficient information to solve this hard task. Users' task performance improves significantly over the baseline representations if they are given our summaries, and gets almost as good as their task performance with the full text.

2 Relations between Papers

Relations between scientific papers, and in particular relations between previous approaches in the literature, are crucial to researchers (Shum, 1998). Concrete information needs involving scientific relations might occur when writing a paper, when the researcher needs to flesh out an argument; vague information needs involving relations might occur when a researcher new to a field requires an overview of existing approaches and their relation.

In particular, a researcher might want to know about criticisms of prior approaches. Which approach is criticising a given work? What are rival approaches to a given work? How can the contrast between similar approaches best be characterized?

Similarly, one might want to know which research builds on which other work. Papers typically do not describe inventions which are entirely new; instead research builds on prior work, either by the same authors, or by the same *school of thought.* In the field of computational linguistics, the use of the same grammar formalism or statistical framework can constitute such continuities. Also, one piece of research can incorporate parts of a solution from previous work, e.g., by using somebody else's tools, data or mathematical formulæ.

Formal citations are an important indicator of relations between articles. Citation behaviour often shows which researchers are in the same school of thought, as such researchers tend to cite each other more often, and in a more positive way. Recently, citation induction tools have emerged which can automatically create citation-indexes of full-text papers, e.g. work by Nanba and Okumura (1999), or Lawrence et al.'s (1999) CiteSeer.

If citation information were combined with information about the type of relations between papers (contrastive and supportive), a sophisticated IR environment could be designed specifically for the information needs of researchers. Such a system could support queries like "*Which approaches are mentioned in the papers about pronoun resolution?*" and "*Of these papers, show me all which use or build on Centering Theory.*" Alternatively, the researcher could ask for approaches *criticising* Centering Theory.

Our summaries, which are specialized in describing the goal of a paper in relation to other work, should be seen in this context. They consist of

- sentences describing the goal of the paper (AIM);
- sentences describing which other approaches are criticised (CONTRAST); and

- sentences describing which other approaches contribute a part of the solution (Basis).

In contrast to previous sentence extracts, e.g. the ones examined by Mani et al. (1999a) and Tombros et al. (1998), ours show a much higher compression: 5% of the full text, as opposed to 17% and 15%. This high compression is necessitated by the genre we work with, scientific texts. Such texts are typically much longer than news wire text.

3 The Data

Our corpus consists of 80 conference articles in the field of computational linguistics, collected from the computation and language archive (CMP_LG, 1994). We chose papers from major conferences and associated workshops; nevertheless we noticed a high level of variability in our corpus, with respect to subtopic within computational linguistics, writing style, quality of English and presentational tradition.

The papers, initially in LaTeX format, are processed automatically with our implementation which uses the TTT system (Grover et al., 1999), such that paragraphs, headlines, section structure, formal citations, sentence borders, and POS information for each word are determined and encoded in XML. Mathematical equations are replaced by placeholders.

We previously collected human judgements about the rhetorical status of each sentence for a set of 25 articles (Teufel et al., 1999). The definition of rhetorical status we use is given in Figure 1.

The subjects classified each sentence into these seven, mutually exclusive categories ("rhetorical contexts"). Written guidelines (17 pages) give strategies for dealing with conflicts between assignments of labels. Three task-trained human annotators reached an inter-annotator agreement of K=.71 (N=4261, k=3[1]). This level of agreement is generally agreed as very reliable annotation. Considering agreement for the categories which interest us here, we observed precision and recall values (measured between two annotators, if one is taken as the gold standard) as follows: 72% precision/56% recall (Aim); 50% precision/55% recall (Contrast) and 82% precision/34% recall (Basis).

Aim	Specific research goal of the current paper
Textual	Statements about section structure
Own	(Neutral) description of own work presented in current paper: Methodology, results, discussion
Background	Generally accepted scientific background
Contrast	Statements of comparison with or contrast to other work; weaknesses of other work
Basis	Statements of agreement with other work or continuation of other work
Other	(Neutral) description of other researchers' work

Figure 1: Annotation Scheme for Rhetorical Status

Absolute Location in Paper
Relative Location of Sentence within Section
Relative Location of Sentence within Paragraph
Type of Headline of Current Section
Sentence Length
Presence of TF*IDF Words in Sentence
Presence of Title Words in Sentence
Voice of First Verb in Sentence
Tense of First Verb in Sentence
Presence of Modal Auxiliary
Presence, Location and Type of Citation
Presence and Type of Formulaic Expression
Presence and Type of Agent
Presence and Type of Action, Presence of Negation
Category of Previous Sentence

Figure 2: Features Used in System

The human annotation of the entire development corpus (80 papers, annotated by one judge) is used as training material for our system. We also use the human annotation of Aim, Contrast and Basis sentences as one of the document representations given to subjects in our evaluation (see section 5).

4 The System

Our system uses machine learning to relate objectively identifiable features of each sentence in text (e.g. the number and location of citations occurring in that sentence) with a human-assigned rhetorical label for that sentence (e.g. Aim or Contrast).

Figure 2 summarises our features, some of which are introduced by us, whereas others stem from previous sentence extraction work (Luhn, 1958; Baxendale, 1958; Paice, 1990). The features cover different aspects of document structure, such as the grammatical subject of a sentences (i.e., the authors or other researchers), the types of actions reported, the location of a sentence, and the presence of citations. The Agent feature, for instance, contains patterns covering pronouns and researchers' proper names.

We use a Naive Bayesian classifier first introduced by Kupiec et al. (1995) and train it on our development corpus of 80 papers. Given unseen text, the

[1] K stands for the Kappa coefficient (Siegel and Castellan, 1988), N for the number of items (sentences) annotated and k for the number of annotators.

AIM

22 We now give a similarity-based method for estimating the probabilities of cooccurrences unseen in training.

151 Our method combines similarity-based estimates with Katz's back-off scheme, which is widely used for language modeling in speech recognition. (BASIS)

CONTRAST

20 Their model, however, is not probabilistic, that is, it does not provide a probability estimate for unobserved cooccurrences.

28 We applied our method to estimate unseen bigram probabilities for Wall Street Journal text and compared it to the standard back-off model. (OWN)

115 We will outline here the main parallels and differences between our method and cooccurrence smoothing.

BASIS

23 Similarity-based estimation was first used for language modeling in the cooccurrence smoothing method of Essen and Steinbiss (1992), derived from work on acoustic model smoothing by Sugawara et al. (1985). (OTHER)

87 The baseline back-off model follows closely the Katz design, except that for compactness all frequency one bigrams are ignored.

122 Notice that this formula has the same form as our similarity model <CREF/>, except that it uses confusion probabilities where we use normalized weights. (CONTRAST)

Figure 3: Sample System Output (Condition **S**)

features of each sentence are automatically identified, and the classifier returns a category for the sentence. Details of the features or the classifier are given in Teufel and Moens (2000).

Previous cross-validation evaluation measured the annotation accuracy (between one annotator and the system) at K=.45 (N=12484, k=2). Precision and recall values (taking this annotator as gold standard) per category were as follows: 44% precision/65% recall (AIM); 34% precision/20% recall (CONTRAST) and 37% precision/40% recall (BASIS). Compared to different baselines established by (a) a TF*IDF text classifier, (b) random or (c) most-frequent-category, these results represent a considerable improvement. However, they remain significantly below human performance. This cross-evaluation constitutes an *intrinsic* evaluation of the system output (by comparison to a gold standard). Part of the contribution of this paper is to provide an *extrinsic* evaluation of the quality of the system output.

Figure 3 exemplifies the system output for one of the papers used in the experiment, namely "Similarity-Based Estimation of Word Cooccurrence Probabilities" by Dagan, Pereira and Lee (ACL 1993; cmp_lg/9405001). It shows the two AIM sentences that the system found, and three sentences for both category CONTRAST and BASIS, along with their sentence numbers.[2] In case of system misclassification with respect to the judge's decision, the judge's decision is given in parentheses.

We now turn to the extrinsic evaluation itself.

5 Extrinsic Evaluation

5.1 Experimental Design

The experiment is a six-by-six cross design (6 groups and 6 conditions), cf. Figure 4. The cross design is necessary due to the high level of variation between the items (the papers). As it is hard to control for the variation in the quality of the papers (or the quality of their document representations), the normal remedy for the variation would be to raise the number of items a subject sees. This was not an option, as we designed the experiment so that it can be performed in less than one hour per subject. Instead, we created six different experimental groups who are shown randomly-selected papers in different conditions, whereby each subject is his or her own control with respect to other conditions. This design should factor out the difference between papers (and between subjects).

	Papers					
	1	2	3	4	5	6
Group 1	F	A	G	R	K	S
Group 2	A	G	R	K	S	F
Group 3	G	R	K	S	F	A
Group 4	R	K	S	F	A	G
Group 5	K	S	F	A	G	R
Group 6	S	F	A	G	R	K

Figure 4: Experimental Cross Design

5.1.1 Conditions

The six representations (conditions) are:

[2] The 3 BASIS and CONTRAST sentences were random sampled from the full system output, cf. section 5.

F: The full article, presented in printed form
A: the author-written abstract
K: a list of keywords, as derived by a TF*IDF measure
R: a random selection of sentences from the abstract
S: Aim, Basis and Contrast sentences, as generated by our system
G: Aim, Basis and Contrast sentences, gained from human annotation as described in section 3 (gold standard)

1. *What is the goal/contribution of the paper?*
2. *Contrastive approaches*
   (a) Which approaches are mentioned? Identify them by citation or informal name.
   (b) What is the criticism/ difference/ contrast?
3. *Prior approaches which are part of the solution*
   (a) Which other approaches are mentioned?
   (b) In which respect is their solution included?
4. *How useful did you find the information you were given to solve the task? Indicate on a scale from 10 to 1, with 10 being extremely useful and 1 being useless.*
5. *Did you know this paper beforehand? Is this paper closely connected to your own research or field of expertise?*

Figure 5: Questions Asked in Task-Based Evaluation

The design of this experiment answers two questions: First, the difference between Conditions **F, A, K, R** and **G** tells us which representation is well-suited to the task of describing relations between papers, and how well and consistently humans can perform the task. Second, the difference between Conditions **S** and **G** tells us how well the system output approximates the gold standard.

5.1.2 Materials

Six papers were randomly chosen from the 80 papers of our development corpus, and document representations in all six conditions were created for them.

We controlled the length of representations in Conditions **K, R, S** and **G**. Conditions **S** and **G** have at most three sentences for each of the categories Aim, Contrast and Basis. If the system or human-decided gold standard reported more than three sentences per condition, three were chosen at random. This was done because we wanted to keep the amount of information presented across items constant, but papers contain a different amount of Aim, Basis, or Contrast sentences. The random choice of sentences for condition **R** was constrained in such a way that the number of words in the selected sentences approximately matched the number of words occurring in Condition **G**. Condition **K** presents as many keywords as there are noun phrases in Condition **G**. As a result, the amount of information presented in conditions **S, G, R** and **K** is kept approximately the same.

5.1.3 Subjects

24 subjects participated. 21 are graduate students and faculty working in computational linguistics (Columbia University and Edinburgh University), 3 are graduate students in other fields of computer science from Columbia. Each experimental group consists of 4 subjects. Not all subjects were native speakers of English, but all can be expected to be familiar with the field of computational linguistics, and accustomed to extracting information from scientific articles.

5.1.4 Procedure

Each experimental group sees each of the six papers in a different representation (condition), but in the same order of articles. Subjects are also given the title of the paper. Skim-reading time of the full text condition was restricted to 10 minutes. After presentation of each paper, the subject was asked to answer the five questions explained below. While filling in the answer-sheet, the subjects also had access to the citation list of the paper. Task completion time, though not formally measured, was much lower in the document surrogate conditions **A, K, R, S, T** than it was in the full paper condition (**F**). Total task completion time was around 40–50 minutes for all six conditions.

5.1.5 Questions

Each subjects answers the 5 questions given in Figure 5 about each paper. Questions 1, 2, and 3 measure task performance, and produce *Task Scores* (TS) which are scored manually. Question 4 measures task adequacy in a subjective way, producing a measure we call the *Utility Score*. This score, ranging from 1 (useless) to 10 (very useful), is interpreted as a measure of the subjects' confidence in their task performance.

Question 5 was used because we use subjects of different levels of expertise. In a later analysis, we might decide to rule out researchers who knew the papers too well; in this analysis, this information was not used, however.

Answers are collected by asking subjects to fill in a tabular answer sheet. As an example, Figure 6 shows the answers of one subject in condition **S** after they saw the information in Figure 3.

5.1.6 Scoring the Task Performance

We scored subjects' answers by comparing each approach the subjects listed with a gold standard, i.e. a definition of the right answer. The gold standard is defined as the combined answers of those four subjects in the group that saw the full papers (Condition

1. Aim: Extending co-occurrence probabilites of unseen events using similarity messures and a corpus

2. Contrast:

(a) Approach	**(b) Relation**
?	not probabilistic
cooccurrence smoothing (Essen, Steinbiss, 92)	differences
Katz 1987 standard back-off model	differences

3. Basis:

(a) Approach	**(b) Relation**
Katz 1987 back-off model	further development
Essen & Steinbiss 92	idea and formula

4. Usefulness: 6

Figure 6: An Example of an Answer Sheet

Contrasted Approaches	**Weight**
Essen and Steinbiss (1992)	3
Brown et al. (1992), class-based models	2
Dagan et al. (1993)	1
Grishman and Sterling (1993)	1
Katz (1987)	1
	8

Supported Approaches	**Weight**
Katz (1987)	3
Pereira et al. (1993)	3
Paul (1991)	1
Dagan et. al (1993)	1
Essen and Steinbiss (1992)	1
Baseline bigram model (MIT)	1
	10

Figure 7: Gold Standard Answers

F). These subjects who saw the full papers arguably had access to the "maximum" available information.

The gold standard also assigns a weight to each approach, which is defined as the number of judges who identified the given approach (thus ranging between 1 and 4). As we assume that more *prominent* approaches are noticed by more judges, the weight should reflect the relevance of an approach. Figure 7 shows the gold standard and the weights for the example paper.

The final score is normalized to 1 by dividing by the sum of all weights for this question and paper. This way of scoring has the positive effect that each paper contributes the same amount to the final score.[3]

There is much more variation in the answers with respect to relations (right side of Figure 6) than there is with respect to approaches (left side). Counting approaches gives a simple measure of how much a subject understood about the relations to cited papers, without having to subjectively judge the depth of understanding in each answer. However, approaches mentioned without any relation were discarded – in order to score points, the subjects should have understood at least *something* about each approach they listed.

Each answer sheet was scored by assigning the corresponding weight from the gold standards. In rare cases, when it was not obvious if two accounts should be counted as identical, half the score was assigned (e.g. if a description of an approach was generally correct, but considered as too vague). The answer sheet in Figure 6, for instance, scored 0.5+3+1=4.5[4] out of 8 for CONTRAST and 1+3 out of 10 for BASIS, and 8.5 out of 18 for the Combined Task Score (CONTRAST and BASIS combined).

During scoring, it turned out that it was substantially harder to determine Task Score AIM (question 1) than Task Scores CONTRAST and BASIS (questions 2 and 3). Most subjects, having read the title, could guess the goal of the paper more or less well. Only in four out of the 24 x 6 = 144 single conditions was a subject unable to guess the aim of the paper. Instead, the answers differed in depth of understanding and specificity, and we found ourselves unable to judge the quality of the answers objectively. We therefore do not report on the task score AIM in this paper.

We found hardly any *wrong* approaches in the answers: subjects seem to only have identified approaches if they felt sure that the approach was correct. As a result, precision was 100% in almost all cases. We therefore only report recall. To summarise, the following variables, with answers coming from different questions, are measured:

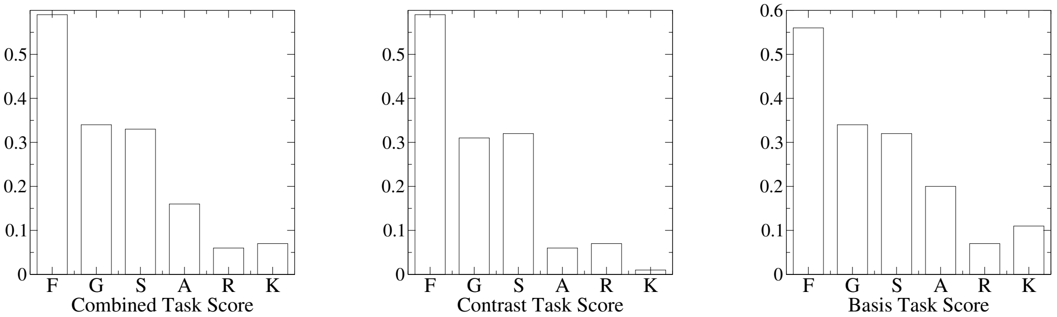
Variable	**Question**	**Scores**
TS CONTRAST	2	[0..1]
TS BASIS	3	[0..1]
Combined TS (BASIS and CONTRAST)	2, 3	[0..1]
Utility Score	4	1, 2,.. 10

5.2 Results

Figure 8 gives the average task scores (recall) for the six conditions. A Wilcoxon matched-pairs signed-

[3] This, however, puts more emphasis on an approach if it appears in a paper mentioning only few other approaches, as opposed to if it appears in a paper mentioning more approaches.

[4] The underspecified reference to Dagan et al.'s approach achieved half the score.



Conditions	TS Combined	TS Contrast	TS Basis
F Full text	0.59	0.59	0.56
G Gold Standard	0.34	0.31	0.34
S System Output	0.33	0.32	0.32
A Abstracts	0.16	0.06	0.20
R Random	0.06	0.07	0.07
K Keywords	0.07	0.01	0.11

Figure 8: Mean Task Scores for the Six Conditions

rank test (Siegel and Castellan, 1988) found all differences between conditions to be statistically significant at p<0.01, except in the following cases:

Combined Task Score:

G and **S**	not significant
K and **R**	not significant
A and **R**	significant at p<.05
A and **K**	significant at p<.05

Contrast Task Score:

G and **S**	not significant
K and **R**	not significant
A and **R**	not significant
A and **K**	not significant

Basis Task Score:

G and **S**	not significant
K and **R**	not significant
A and **K**	not significant
A and **S**	significant at p<.05
A and **G**	significant at p<.02
A and **R**	significant at p<.02

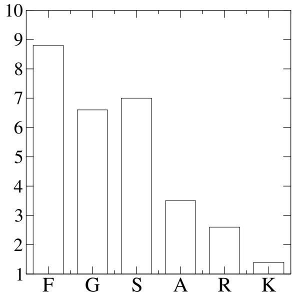
The task scores are our main indication of the task adequacy of the different document surrogates. It is clear that our summary lists, both in the gold standard version (**G**) and as actual system output (**S**), are very well suited to the task (TS of .34 and .33); they provide significantly more information than the abstract (TS of .16), keywords (TS of .06) or random sentences (TS of 0.07). One should also take into account the particularly high compression and the fact that conditions **G** and **S** did not show the full amount of information but were cut to 3 random sentences per category; subjects would most likely have performed better with the full set of information.

Overall, task scores are better for BASIS than they are for CONTRAST. It seems easier to guess from restricted information which school of thought an approach belongs to than it is to guess which specific other approaches are criticised. One reason why this might be the case is that BASIS contributions are often described in one single sentence, but CONTRAST connections can be more complex and might stretch over several sentences. In this case, it is hard for our system (or any sentence-extraction based system) to pick the right sentence.

The comparison of other conditions with condition **F** is slightly problematic as the scores of the other five conditions depend on the majority-style answers given in condition **F**. We can say that task performance within the **F** group was on average 60% of the 100% that are theoretically possible in that condition and group, but numerical comparisons to other conditions are not fully supported with our experimental design. The scoring method also overestimates scores for condition **F**.[5] In an ideal world, we would have an independent source of gold standards, e.g. several annotators which are given indefinite time with the papers and which decide for each citation if it is mentioned in contrastive or supportive context.

Figure 9 shows that the subjective Utility Scores

[5] For instance, correct relations might have been overlooked by the **F** subjects in the 10 minute skim-read, whereas other subjects might have detected these in other conditions. In this case, the gold standard does not punish the **F** subjects' oversight.



Conditions	Utility Score
F Full text	8.8
G Gold Standard	6.6
S System Output	7.0
A Abstracts	3.5
R Random	2.6
K Keywords	1.4

Figure 9: Utility Score

generally mirror the task performance values discussed above. A Wilcoxon matched-pairs signed-rank test found all differences to be statistically significant at p<0.01, except in the following cases:

G and **S**	not significant
R and **K**	not significant
A and **K**	significant at p<.05
A and **R**	significant at p<.05

That means that subjects were aware of the suitability of different document surrogates for the task. In general, they were very satisfied with the summary-lists produced by our system. Indeed, many subjects informally remarked how much work it was to extract the approaches from the full text, and how convenient conditions **G** and **S** were (provided that the information in them was reliable).

Human-written generic abstracts are document surrogates which are at a disadvantage in our experimental setup. They generally do not discuss relations to other approaches; they were not written to support our task. Not surprisingly, the Utility Score also shows that subjects did not judge abstract information as useful. A similar picture emerges from the task scores: abstracts perform badly in the Contrast task, while they prove to be more task-adequate in the combined and Basis task.

Keywords or random sentences are not at all useful for the task, as confirmed by both task and utility scores. In general, one would assume that random sentences should do better than keywords because useful sentences might have been selected at random. This effect is present for Task Score Contrast and for the Utility Score; however, the results for Combined Task Score and for Task Score Basis show the reverse effect. We think that this is due to the fact that one well-known paper happened to occur in a keyword condition with three well-read subjects who could guess which work this work was based on (but not which particular work was criticised). In other words, the artificially good performance of keywords for the Basis task is likely to be noise, which would be eliminated if more data was available.

6 Limitations and Future Work

Several things could be improved in this experimental design in the future.

One question is the level of expertise of the subjects, which is crucial. Subjects who are too well-informed might have prior knowledge of the papers, in which case they are likely to perform reasonably well even in the less informative conditions. Subjects who are not well-informed enough might decrease the quality of the gold standard. Ideally, the subjects used in this experiment should be semi-expert subjects at the same level of expertise, as our final system is aimed at semi-experts in the field, who need information about current approaches, rival approaches and continuation relationships. But subjects are hard to come by. Our experimental design at least makes sure that the scores of each subject are countered by their own scores in the other conditions.

Another factor is paper quality, which is hard to control for in a sensible way, other than use respected data sources when compiling the corpus, which we did. Additionally, one might use information about the *impact* of a paper, e.g. as meassured by subsequent citations to that paper.

Our baselines are the type of document surrogates normally encountered in information seeking tasks, but there are harder baselines we will consider in future work. For instance, one condition could have presented randomly sampled sentences containing citations, or sentences randomly sampled from the *related work* section.

7 Discussion

We have proposed a new task for extrinsic evaluation here for the first time. We have to ask ourselves how natural this task is: is it the kind of task people do during their daily work? Information analysts, who have to decide under time pressure which papers to read, routinely perform relevance decisions. We believe that the task of assessing relations between articles is an important task in the daily life of a researcher (e.g. "Has anybody applied this approach recently?"). But this is less obvious than in the case of the analyst because the effects of the researchers' information-foraging are less systematic and less observable from the outside. Indeed, it might be the case that the "harder" tasks, those which are capable of proving the added value of summaries, are inherently less well-defined than simpler tasks. In the end, only the final application—a system allowing researchers to search rhetorical citation relations—will answer the question in how far citation relations are of practical interest for researchers.

Our experiment raises questions about the status of intrinsic versus extrinsic evaluation. A previous intrinsic evaluation (Teufel and Moens, 2000) reported rather low values for the direct comparison of the system's output with the "ideal" annotation the system was trained on: the overall annotation similarity was K=.45; precision ranged between 44% and 34% and recall between 65% and 20% for the categories AIM, BASIS and CONTRAST. However, the present *extrinsic* evaluation showed *no* statistically significant difference in task performance between gold standards and system output – humans could solve the same task equally well with either, and much better than with typical baselines.[6] This means, foremost and all, that evaluation by "ideal summaries", while useful for system tuning, should not be used as final evaluation of a system. As many "good" summaries are possible, comparison of a system's output to *one* such summary will invariably give distorted results.

Another point concerns subjective evaluation scores, like our Utility Score. There are many traditional arguments against using such scores: 1. subjects (who might know the experimentator's work) might guess which summary is produced by the system to be evaluated and be unduly biased in their judgement[7]; 2. if the concept to be evaluated ("utility", "coherence") is not well-defined, judges might use their own idiosyncratic definitions, which makes it hard to meaningfully compare the numerical values; 3. the same might happen if the single scores are not well-defined ("1 means the sentence can be understood, but is barely grammatical").

All these points are true, but in the end it is the user's satisfaction which constitutes the ultimate evaluation—the best summarisation system is the one whose summaries the users want to use most. Therefore, a mixture of task-based evaluation, subjective evaluation and evaluation by "ideal summary" seems the best option for now.

8 Conclusion

In this paper, we have presented a novel method for extrinsic evaluation of summaries, which is based on questions about relations between scientific approaches. In contrast to the task of relevance decision, this task is "harder" in that mere information about a paper's topic will not help. Indeed, the task is designed in such a way that the added value of summaries can be shown in comparison to simpler document representations, such as a list of keywords or random sentences.

The experiment is extremely time-efficient. Each subject produces several task-scores for the 6 conditions he or she sees within 40–50 minutes, far more data points than a relevance decision task would produce in the same time. This makes it feasible to test against multiple baselines in one experiment. The preparation of the materials only requires the generation of the different document surrogates, the preparation of the gold standard and the final scoring of the answers. In contrast, material preparation for relevance decision tasks is notoriously time-consuming. Using and IR system, queries need to be found which are not too specific and not too inspecific. All returned documents must be judged by a human into relevant or irrelevant in order to be able to calculate precision and recall, a task that can take several hours if the document set is reasonably large.

We have established the following results about the usefulness of different types of document representations for the task of defining relations between papers: Keyword lists and random sentences do not provide enough information to enable subjects to describe relations between papers. Author-provided generic abstracts also do not provide enough information for the task of describing contrasted approaches, while they provide adequate information for the task of describing supportive approaches. Summary-lists produced by our system (Teufel and Moens, 2000), provide enough (and the right kind of) information to do the task, as do full papers (the ceiling condition in this experiment).

We also experimentally established that the output of our system was not statistically significantly different from the gold standard for the task. These positive results were corroborated by users' judgements of usefulness.

References

Barzilay, Regina, Kathleen R. McKeown, and Michael Elhadad. 1999. Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 550–557.

Baxendale, Phyllis B. 1958. Man-Made Index for Technical Literature—an Experiment. *IBM Journal of Research and Development* 2(4): 354–361.

Bazerman, Charles. 1985. Physicists Reading Physics, Schema-Laden Purposes and Purpose-Laden Schema. *Written Communication* 2(1): 3–23.

[6] This could be due to the redundancy in the materials: one mention of an approach was enough for the humans to do the task.

[7] In contrast, it is very hard to "cheat" or do the experimentators a favour in a task-based evaluation.

CMP_LG. 1994. The Computation and Language E-Print Archive, `http://xxx.lanl.gov/cmp-lg`.

Grefenstette, Gregory. 1998. Producing Intelligent Telegraphic Text Reduction to Provide an Audio Scanning Service for the Blind. In Radev and Hovy 1998, 111–117.

Grover, Claire, Andrei Mikheev, and Colin Matheson. 1999. LT TTT Version 1.0: Text Tokenisation Software. Technical report, Human Communication Research Centre, University of Edinburgh. `http://www.ltg.ed.ac.uk/software/ttt/`.

Jing, Hongyan, and Kathleen R. McKeown. 2000. Cut and Paste Based Summarization. In *Proceedings of the 6th Applied Natural Language Processing(ANLP-00) and the 1st North American Chapter of the Association of Computational Linguistics (NAACL-00)*, 178–185.

Knight, Kevin, and Daniel Marcu. 2000. Statistics-Based Summarization — Step One: Sentence Compression. In *Proceeding of The 17th National Conference of the American Association for Artificial Intelligence (AAAI-2000)*, 703–710.

Kupiec, Julian, Jan O. Pedersen, and Francine Chen. 1995. A Trainable Document Summarizer. In *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR-95)*, 68–73.

Lawrence, Steve, C. Lee Giles, and Kurt Bollacker. 1999. Digital Libraries and Autonomous Citation Indexing. *IEEE Computer* 32(6): 67–71.

Luhn, Hans Peter. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* 2(2): 159–165.

Mani, Inderjeet, Therese Firmin, David House, Gary Klein, Beth Sundheim, and Lynette Hirschman. 1999a. The TIPSTER Summac Text Summarization Evaluation. In *Proceedings of the 9th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99)*, 77–85.

Mani, Inderjeet, Barbara Gates, and Eric Bloedorn. 1999b. Improving Summaries by Revising them. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, 558–565.

Nanba, Hidetsugu, and Manabu Okumura. 1999. Towards Multi-Paper Summarization using Reference Information. In *Proceedings of IJCAI-99*, 926–931.

Paice, Chris D. 1990. Constructing Literature Abstracts by Computer: Techniques and Prospects. *Information Processing and Management* 26: 171–186.

Radev, Dragomir R., and Eduard H. Hovy, eds. 1998. *Working Notes of the AAAI Spring Symposium on Intelligent Text Summarization*.

Shum, Simon Buckingham. 1998. Evolving the Web for Scientific Knowledge: First Steps towards an "HCI Knowledge Web". *Interfaces, British HCI Group Magazine* 39: 16–21.

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

Teufel, Simone, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the 8th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-99)*, 110–117.

Teufel, Simone, and Marc Moens. 2000. What's yours and what's mine: Determining Intellectual Attribution in Scientific Text. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Tombros, Anastasios, Mark Sanderson, and Phil Gray. 1998. Advantages of Query Biased Summaries in Information Retrieval. In Radev and Hovy 1998.