Sentence extraction as a classification task

Simone Teufel
Centre for Cognitive Science
and Language Technology Group
University of Edinburgh
S.Teufel@ed.ac.uk

Marc Moens
Language Technology Group
University of Edinburgh
M.Moens@ed.ac.uk

Abstract

A useful first step in document summarisation is the selection of a small number of 'meaningful' sentences from a larger text. Kupiec et al. (1995) describe this as a classification task: on the basis of a corpus of technical papers with summaries written by professional abstractors, their system identifies those sentences in the text which also occur in the summary, and then acquires a model of the 'abstract-worthiness' of a sentence as a combination of a limited number of properties of that sentence.

We report on a replication of this experiment with different data: summaries for our documents were not written by professional abstractors, but by the authors themselves. This produced fewer alignable sentences to train on. We use alternative 'meaningful' sentences (selected by a human judge) as training and evaluation material, because this has advantages for the subsequent automatic generation of more flexible abstracts. We quantitatively compare the two different strategies for training and evaluation (viz. alignment vs. human judgement); we also discuss qualitative differences and consequences for the generation of abstracts.

1 Introduction

A useful first step in the automatic or semi-automatic generation of abstracts from source texts is the selection of a small number of 'meaningful' sentences from the source text. To achieve this, each sentence in the source text is scored according to some measure of importance, and the best-rated sentences are selected. This results in collections of the N most 'meaningful' sentences, in the order in which they appeared in the source text – we will call these *excerpts*. An excerpt can be used to give readers an idea of what the longer text is about, or it can be used as input into a process to produce a more coherent abstract.

It has been argued for almost 40 years that it is possible to automatically create excerpts which meet basic information compression needs (Luhn, 1958). Since then, different measurements for the importance of a sentence have been suggested, in particular stochastic measurements for the significance of key words or phrases (Luhn, 1958; Zechner, 1995). Other research, starting with (Edmundson, 1969), stressed the importance of heuristics for the location of the candidate sentence in the source text (Baxendale, 1958) and for the occurrence of cue phrases (Paice and Jones, 1993; Johnson et al., 1993).

Single heuristics tend to work well on documents that resemble each other in style and content. For the more robust creation of excerpts, combinations of these heuristics can be used. The crucial question is how to combine the different heuristics. In the past, the relative usefulness of single methods had to be balanced manually. Kupiec et al. (1995) use supervised learning to automatically adjust feature weights, using a corpus of research papers and corresponding summaries.

Humans have good intuition about what makes a sentence 'abstract-worthy', i.e. suitable for inclusion in a summary. Abstract-worthiness is a high-level quality, comprising notions such as semantic content, relative importance and appropriateness for representing the contents of a document. For the automatic evaluation of the quality of machine generated excerpts, one has to find an operational approximation to this subjective notion of abstract-worthiness, i.e. a definition of a desired result. We will call the criteria of what constitutes success the *gold standard*, and the set of sentences that fulfill

these criteria the *gold standard sentences*. Apart from evaluation, a gold standard is also needed for supervised learning.

In Kupiec et al. (1995), a gold standard sentence is a sentence in the source text that is matched with a summary sentence on the basis of semantic and syntactic similarity. In their corpus of 188 engineering papers with summaries written by professional abstractors, 79% of sentences occurred in both summary and source text with at most minor modifications.

However, our collection of papers, whose abstracts were written by the authors themselves, shows a significant difference: these abstracts have significantly fewer alignable sentences (31.7%). This does not mean that there are fewer abstract-worthy sentences in the source text. We used a simple (labour-intensive) way of defining this alternative gold standard, viz. asking a human judge to identify additional abstract-worthy sentences in the source text.

Our main question was whether Kupiec et al.'s methodology could be used for our kind of gold standard sentences also, and if there was a fundamental difference in extraction performance between sentences in both gold standards or between documents with higher or lower alignment. We also conducted an experiment to see how additional training material would influence the statistical model.

The remainder of this paper is organized as follows: in the next section, we summarize Kupiec et al.'s method and results. Then, we describe our data and discuss the results from three experiments with different evaluation strategies and training material. Differences between our and Kupiec et al.'s data with respect to the alignability of document and summary sentences, and consequences thereof are considered in the discussion.

2 Sentence selection as classification

In Kupiec et al.'s experiment, the gold standard sentences are those summary sentences that can be aligned with sentences in the source texts. Once the alignment has been carried out, the system tries to determine the characteristic properties of aligned sentences according to a number of features, viz. presence of particular cue phrases, location in the text, sentence length, occurrence of thematic words, and occurrence of proper names. Each document sentence receives scores for each of the features, resulting in an estimate for the sentence's probability to also occur in the summary. This probability is calculated as follows:

$$P(s \in S|F_1,\ldots,F_k) \approx \frac{P(s \in S)\prod_{j=1}^{k} P(F_j|s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

$P(s \in S\|F_1,\ldots,F_k)$:	Probability that sentence s in the source text is included in summary S, given its feature values;
$P(s \in S)$:	compression rate (constant);
$P(F_j\| s \in S)$:	probability of feature-value pair occurring in a sentence which is in the summary;
$P(F_j)$:	probability that the feature-value pair occurs unconditionally;
k:	number of feature-value pairs;
F_j:	j-th feature-value pair.

Assuming statistical independence of the features, $P(F_j|s \in S)$ and $P(F_j)$ can be estimated from the corpus.

Evaluation relies on cross-validation. The model is trained on a training set of documents, leaving one document out at a time (the current test document). The model is then used to extract candidate sentences from the test document, allowing evaluation of precision (sentences selected correctly over total number of sentences selected) and recall (sentences selected correctly over alignable sentences in summary). Since from any given test text as many sentences are selected as there are alignable sentences in the summary, precision and recall are always the same.

Kupiec et al. reports that precision of the individual heuristics ranges between 20–33%; the highest cumulative result (44%) was achieved using paragraph, fixed phrases and length cut-off features.

3 Our experiment

3.1 Data and gold standards

Our corpus is a collection of 202 papers from different areas of computational linguistics, with summaries written by the authors.[1] The average length of the summaries is 4.7 sentences; the average length of the documents 210 sentences.

We semi-automatically marked up the following structural information: title, summary, headings, paragraph structure and sentences. Tables, equations, figures, captions, references and cross references were removed and replaced by place holders.

[1] The corpus was drawn from the computation and language archive (`http://xxx.lanl.gov/cmp-lg`), converted from LaTeX source into HTML in order to extract raw text and minimal structure automatically, then transformed into our SGML format with a perl script, and manually corrected. Data collection took place collaboratively with Byron Georgantopolous.

Figure 1: Composition of gold standards for training sets

We decided to use two gold standards:

- **Gold standard A: Alignment.** Gold standard sentences are those occurring in both author summary and source text, in line with Kupiec et al.'s gold standard.
- **Gold standard B: Human Judgement.** Gold standard sentences are non-alignable source text sentences which a human judge identified as relevant, i.e. indicative of the contents of the source text. Exactly how many human-selected sentence candidates were chosen was the human judge's decision.

Alignment between summary and document sentences was assisted by a simple surface similarity measure (longest common subsequence of non-stoplist words). Final alignment was decided by a human judge. The criterion was similarity of semantic contents of the compared sentences. The following sentence pair illustrates a *direct match*:

> **Summary:** *In understanding a reference, an agent determines his confidence in its adequacy as a means of identifying the referent.*
>
> **Document:** *An agent understands a reference once he is confident in the adequacy of its (inferred) plan as a means of identifying the referent.*

Our data show an important difference with Kupiec et al.'s data: we have significantly lower alignment rates. Only 17.8% of the summary sentences in our corpus could be automatically aligned with a document sentence with a certain degree of reliability, and only 3% of all summary sentences are identical matches with document sentences.

We created three different sets of training material:

- **Training set 1:** The 40 documents with the highest rate of overlap; 84% of the summary sentences could be semi-automatically aligned with a document sentence.
- **Training set 2:** 42 documents from the year 1994 were arbitrarily chosen out of the remaining 163 documents and semi-automatically aligned. They showed a much lower rate of overlap; only 36% of summary sentences could be mapped into a document sentence.
- **Training set 3:** 42 documents from the year 1995 were arbitrarily chosen out of the remaining documents and semi-automatically aligned. Again, the overlap was rather low: 42%.
- **Training set 123:** Conjunction of training sets 1, 2 and 3. The average document length is 194 sentences; the average summary length is 4.7 sentences.

A human judge provided a mark-up of additional abstract-worthy sentences for these 3 training sets (124 documents). The remaining 78 documents remain as unseen test data. Figure 1 shows the composition of gold standards for our training sets. Gold standard sentences for training set 1 consist of an approximately balanced mixture of aligned and human-selected candidates, whereas training set 2 contains three times as many human-selected as aligned gold standard sentences, training set 3 even four times as many. Each document in training set 1 is associated with an average of 7.75 gold standard sentences (A+B), compared to an average of 7.07 gold standard sentences in training set 2, and an average of 9.14 gold standard sentences in training set 3.

3.2 Heuristics

We employed 5 different heuristics: 4 of the methods used by Kupiec et al. (1995), viz. cue phrase method, location method, sentence length method and thematic word method, and another well-known method in the literature, viz. title method.

1. Cue phrase method: The cue phrase method seeks to filter out meta-discourse from subject matter. We advocate the cue phrase method as our main method because of the additional 'rhetorical' context these meta-linguistic markers make available. This context of the extracted sentences – along with their propositional content – can be used to generate more flexible abstracts.

We use a list of 1670 negative and positive cues and indicator phrases or formulaic expressions, 707 of which occur in our training sets. For simplicity and efficiency, these cue phrases are fixed strings.

Our cue phrase list was manually created by a cycle of inspection of extracted sentences, identification of as yet unaccounted-for expressions, addition of these expressions to the cue phrase list, and possibly inclusion of overlooked abstract-worthy sentences in the gold standard. Cue phrases were manually classified into 5 classes, which we expected to correspond to the likelihood of a sentence containing the given cue to be included in the summary: a score of -1 means 'very unlikely'; $+3$ means 'very likely to be included in a summary'.[2] We found it useful to assist the decision process with corpus frequencies. For each cue phrase, we compiled its relative frequency in the gold standard sentences and in the overall corpus. If a cue phrase proved general (i.e. it had a high relative corpus frequency) and distinctive (i.e. it had a high frequency within the gold standard sentences), we gave it a high score, and included other phrases that are syntactically and semantically similar to it into the cue list. We scanned the data and found the following tendencies:

- Certain communicative verbs are typically used to describe the overall goals; they occur frequently in the gold-standard sentences (*argue*, *propose*, *develop* and *attempt*). Others are predominantly used for describing communicative sub-goals (detailed steps and sub-arguments) and should therefore be in a different equivalence class (*prove*, *show* and *conclude*). Within the class of communicative verbs, tense and mode seem to be relevant for abstract-worthiness. Verbs in past tense or present perfect (as used in the conclusion) are more likely to refer to global achievements/goals, and thus to be included in the summary. In the body of the text, present and future forms tend to be used to introduce subtasks.

- Genre specific nominal phrases like *this paper* are more distinctive when they occur at the beginning of the sentence (as an approximation to subject/topic position) than their non-subject counterparts.

- Explicit summarisation markers like *in sum*, *concluding* did occur frequently, but quite unexpectedly almost always in combination with communicative sub-tasks. They were therefore less useful at signalling abstract-worthy material.

Sentences in the source text are matched against expressions in the list. Matching sentences are classified into the corresponding class, and sentences not containing cue phrases are classified as 'neutral' (score 0). Sentences with competing cue phrases are classified as members of the class with the higher numerical score, unless one of the competing classes is negative.

Sentences occurring directly after headings like *Introduction* or *Results* are valuable indicators of the general subject area of papers. Even though one might argue that this property should be handled within the location method, we perceive this information as meta-linguistic (and thus logically belonging to the cue phrase method). Thus, scores for these sentences receive a prior score of $+2$ ('likely to occur in a summary').

In a later section, we show how this method performs on unseen data of the same kind (viz. texts in the genre of computational linguistics research papers of about ∼6–8 pages long). Even though the cue phrase method is well tuned to these data, we are aware that the list of phrases we collected might not generalize to other genres. Some kind of automation seems desirable to assist a possible adaptation.

2. Location method. Paragraphs at the start and end of a document are more likely to contain material that is useful for a summary, as papers are organized hierarchically. Paragraphs are also organized hierarchically, with crucial information at the beginning and the end of paragraphs. Therefore, sentences in document peripheral paragraphs should be good candidates, and even more so if they occur in the periphery of the paragraph.

[2] We experimented with larger and smaller numbers of classes, but obtained best results with the 5-way distinction.

Our algorithm assigns non-zero values only to sentences which are in document peripheral sections; sentences in the middle of the document receive a 0 score. The algorithm is sensitive to prototypical headings (*Introduction*); if such headings cannot be found, it uses a fixed range of paragraphs (first 7 and last 3 paragraphs). Within these document peripheral paragraphs, the values 'i_f' and 'm' (for paragraph initial-or-final and paragraph medial sentences, respectively) are assigned.

3. Sentence Length method. All sentences under a certain length (current threshold: 15 tokens including punctuation) receive a 0 score, all sentences above the threshold a 1 score.

Kupiec et al. mention this method as useful for filtering out captions, titles and headings. In our experiment, this was not necessary as our format encodes headings and titles as such, and captions are removed. As expected, it turns out that the sentence length method is our least effective method.

4. Thematic word method. This method tries to identify key words that are characteristic for the contents of the document. It concentrates on non-stop-list words which occur frequently in the document, but rarely in the overall collection. In theory, sentences containing (clusters of) such thematic words should be characteristic for the document. We use a standard term-frequency*inverse-document-frequency (tf*idf) method:

$$score(w) = f_{loc} * log(\frac{100*N}{f_{glob}})$$

f_{loc}: frequency of word w in document
f_{glob}: number of documents containing word w
N: number of documents in collection

The 10 top-scoring words are chosen as thematic words; sentence scores are then computed as a weighted count of thematic word in sentence, meaned by sentence length. The 40 top-rated sentences get score 1, all others 0.

5. Title method. Words occurring in the title are good candidates for document specific concepts. The title method score of a sentence is the mean frequency of title word occurrences (excluding stop-list words). The 18 top-scoring sentences receive the value 1, all other sentences 0. We also experimented with taking words occurring in all headings into account (these words were scored according to the tf*idf method) but received better results for title words only.

	Indiv.	**Cumul.**
Method 1 (cue)	55.2	55.2
Method 2 (location)	32.1	65.3
Method 3 (length)	28.9	66.3
Method 4 (tf*idf)	17.1	66.5
Method 5 (title)	21.7	68.4
Baseline	28.0	

Figure 2: First experiment: Impact of individual heuristics; training set 123, gold standards A+B

	Seen	**Unseen**
Cue Phrase Method	60.9	54.9
Heuristics Combination	71.6	65.3
Baseline	29.1	

Figure 3: First Experiment: Difference between unseen and seen data; training set 3, gold standards A+B

3.3 Results

Training and evaluation took place as in Kupiec et al.'s experiment. As a baseline we chose sentences from the beginning of the source text, which obtained a recall and precision of 28.0% on training set 123. This from-top baseline (which is also used by Kupiec et al.) is a more conservative baseline than random order: it is more difficult to beat, as prototypical document structure places a high percentage of relevant information in the beginning.

3.3.1 First experiment

Figure 2 summarizes the contribution of the individual methods.[3] Using the cue phrase method (method 1) is clearly the strongest single heuristic. Note that the contribution of a method cannot be judged by the individual precision/recall for that method. For example, the sentence length method (method 3) with a recall and precision over the baseline contributes hardly anything to the end result, whereas the title method (method 5), which is below the baseline if regarded individually, performs much better in combination with methods 1 and 2 than method 3 does (67.3% for heuristics 1, 2 and 5; not to be seen from this table). The reason for this is the relative independence of the methods. If method 5 identifies a successful candidate, it is less likely that this candidate has also been identified by method 1 or 2. Method 4 (tf*idf) decreased results slightly in some of the experiments, but not in the

[3] All figures in tables are precision percentages.

	comb	cue	base
TS 1	66.1	49.0	29.6
TS 2	62.2	54.5	24.9
TS 3	71.6	60.9	29.1
TS 123	68.4	55.2	28.0

Figure 4: First experiment: Baseline, best single heuristic and combination; gold standards A+B

	Evaluation strategy					
	Gold standard A			Gold standard B		
TS	**comb**	**cue**	**base**	**comb**	**cue**	**base**
1	36.9	27.5	21.4	45.3	30.4	10.8
2	25.0	18.4	9.2	53.8	47.9	20.3
3	27.1	13.5	13.5	64.3	54.4	25.7
123	31.6	23.2	16.3	57.2	46.7	20.4

Figure 5: Second experiment: Impact of type of gold standard

experiments with our final/largest training set 123 where it led to a (non-significant) increase.

We also checked how much precision and recall decrease for unseen data. This decrease applies only to the cue phrase method, because the other heuristics are fixed and would not change by seeing more data. After the manual mark-up of gold standard sentences and additions to the cue phrase list for training set 3, we treated training set 3 as if it was unseen: we used only those 1423 cue phrases for extraction that were compiled from training set 1 and 2. A comparison of this 'unseen' result to the end result (Figure 3) shows that our cue phrase list, even though hand-crafted, is robust and general enough for our purposes; it generalizes reasonably well to texts of a similar kind.

Figure 4 shows mean precision and recall for our different training sets for three different extraction methods: a combination of all 5 methods ('comb.'); the best single heuristic ('cue'); and the baseline ('base'). We used both gold standards A+B. These results reconfirm the usefulness of Kupiec et al.'s method of heuristic combination. The method increases precision for the best method by around 20%. It is worth pointing out that this method produces very short excerpts, with compressions as high as 2–5%, and with a precision equal to the recall. Thus this is a different task from producing long excerpts, e.g. with a compression of 25%, as usually reported in the literature. Using this compression, we achieved a recall of 96.0% (gold standard A), 98.0% (gold standard B) and 97.3% (gold standards A+B) for training set 123. For comparison, Kupiec et al. report a 85% recall.

3.3.2 Second experiment

In order to see how the different gold standards contribute to the results, we used only one gold standard (A or B) at a time for training and for extraction. Figure 5 summarizes the results.

Looking at Gold standard A, we see that training set 1 is the only training set which obtains a recall that is comparable to Kupiec et al.'s. Incidentally, training set 1 is also the only training set that is comparable to Kupiec et al.'s data with respect to alignability. The bad performance of training set 2 and 3 under evaluation with gold standard A is not surprising, as there are too few aligned gold standard sentences to train on: 50% of the documents in these training sets contain no or only one aligned sentence.

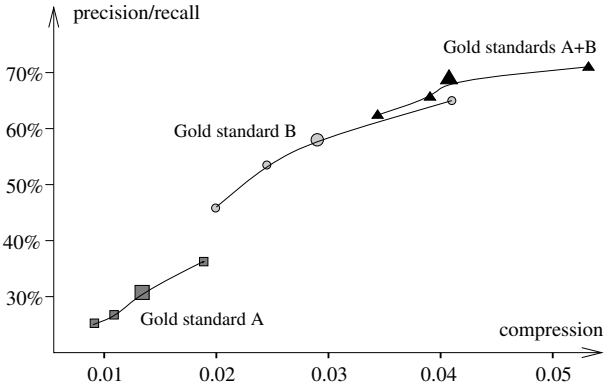

Figure 6: Second experiment: Impact of type of gold standard on precision and recall, as a function of compression

Overall, performance seems to correspond to the ratio of gold standard sentences to source text sentences, i.e. the compression of the task.[4] The dependency between precision/recall and compression is depicted in Figure 6. Taking both gold standards into account increases performance considerably compared to either of the gold standards alone, because of the lower compression. As we don't have training sets with exactly the same number of gold standard A and B sentences, we cannot directly compare the performance, but the graph is suggestive of a similar behaviour of both gold standards. The results for training set 123 fall between the results of the individual training sets (symbolized by the large data points).

[4]The difference in performance between training sets in the first experiment is thus probably mainly attributable to differences in compression between the training sets.

		Extraction			
	TS	**1**	**2**	**3**	**123**
Training	**1**	66.1	61.2	69.7	66.3
	2	65.8	62.2	69.5	66.0
	3	65.1	62.9	71.6	66.1
	123	66.4	62.9	70.8	68.4

Figure 7: Third experiment: Impact of training material on precision and recall; gold standards A+B

From this second experiment we conclude that for our task, there is no difference between gold standard A and B. The crucial factor that precision and recall depends on is the compression of the task.

3.3.3 Third experiment

In order to evaluate the impact of the training material on precision and recall, we computed each possible pair of training and evaluation material (cf. figure 7).

In this experiment, all documents of the training set are used to train the model; this model is then evaluated against each document in the test set, and the mean precision and recall is reported. Importantly, in this experiment none of the other documents in the test set is used for training.

These experiments show a surprising uniformity within test sets: overall extraction results for each training set are very similar. Training on different data does not change the statistical model much. In most cases, extraction for each training set worked best when the model was trained on the training set itself, rather than on more data. Thus, the difference in results between individual training sets is not an effect of data sparseness at the level of heuristics combination.

We conclude from this third experiment that improvement in the overall results can primarily be achieved by improving single heuristics, and not by providing more training data for our simple statistical model.

4 Discussion

Comparing our experiment to Kupiec et al.'s the most obvious difference is the difference in data.

Our texts are likely to be more heterogeneous, coming from areas of computational linguistics with different methodologies and thus having an argumentative, experimental, or implementational orientation. Also, as they are not journal articles, they are not heavily edited. There is also less of a prototypical article structure in computational linguistics than in experimental disciplines like chemical engineering. This makes our texts more difficult to extract from.

The major difference, however, is that we use summaries which are not written by trained abstractors, but by the authors themselves. In only around 20% of documents in our original corpus, sentence selection had been used as a method for summary generation, whereas professional abstractors rely more heavily and systematically on sentences in the source text when creating their abstracts.

Using aligned sentences as gold standard has two main advantages. First, it makes the definition of the gold standard less labour intensive. Second, it provides a higher degree of objectivity. It is a much simpler task for a human judge to decide if two sentences convey the same propositional content, than to decide if a sentence is qualified for inclusion in a summary or not.

However, using alignment as the sole definition for gold standard implies that a sentence is only a good extraction candidate if its equivalent occurs in the summary, an assumption we believe to be too restrictive. Document sentences other than the aligned ones might have been similar in quality to the chosen sentences, but will be trained on as a negative example with Kupiec et al.'s method. Kupiec et al. also recognize that there is not only one optimal excerpt, and mention Rath et al.'s (1961) research which implies that the agreement between human judges is rather low. We argue that it makes sense to complement aligned sentences with manually determined supplementary candidates. This is not solely motivated by the data we work with but also by the fact that we envisage a different task than Kupiec et al. (who use the excerpts as indicative abstracts). We see the extraction of a set of sentences as an intermediate step towards the eventual generation of more flexible and coherent abstracts of variable length. For this task, a whole range of sentences other than just the summary sentences might qualify as good candidates for further processing.[5] One important subgoal is the reconstruction of approximated document structure (cf. rhetorical structure, as defined in RST (Mann et al., 1992)). One of the reasons why we concentrated on cue phrases was that we believe that cue phrases are an obvious and easily accessible source of rhetorical information.

Another important question was if there were other properties following from the main difference between our training sets, alignability. Are documents with a high degree of alignability *inherently*

[5] This is mirrored by the fact that in our gold standards, the number of human-selected sentence candidates outweighed aligned sentences by far.

more suitable for abstraction by our algorithm? It might be suspected that alignability is correlated with a better internal structure of the papers, but our experiments suggest that, for the purpose of sentence extraction, this is either not the case or not relevant. Our results show that our training sets 1, 2 and 3 behave very similarly under evaluation taking aligned gold standards *or* human-selected gold standards into account. The only definite factor influencing the results was the compression rate. With respect to the quality of abstracts, this implies that the strategy which authors use for summary generation – be it sentence selection or complete regeneration of the summary from semantic representation – is a matter of authorial choice and not an indicator of style, text quality, or any aspect that our extraction program is particularly sensitive to. This means that Kupiec et al.'s method of classificatory sentence selection is not restricted to texts which have high-quality summaries created by human abstractors. We claim that adding human-selected gold standards will be useful for generation of more flexible and coherent abstracts, than training on just a fixed number of author-provided summary sentences would allow.

5 Conclusions

We have replicated Kupiec et al.'s experiment for automatic sentence extraction using several independent heuristics and supervised learning. The summaries for our documents were not written by professional abstractors, but by the authors themselves. As a result, our data demonstrated considerably lower overlap between sentences in the summary and sentences in the main text. We used an alternative evaluation that mixed aligned sentences with other good candidates for extraction, as identified by a human judge.

We obtained a 68.4% recall and precision on our text material, compared to a 28.0% baseline and a best individual method of 55.2%. Combining individually weaker methods results in an increase of around 20% of the best method, in line with Kupiec et al.'s results. This shows the usefulness of Kupiec et al.'s methodology for a different type of data and evaluation strategy. We found that there was no difference in performance between our evaluation strategies (alignment or human judgement), apart from external constraints on the task like the compression rate. We also show that increased training did not significantly improve the sentence extraction results, and conclude that there is more room for improvement in the extraction methods themselves.

With respect to our ultimate goal of generating of higher quality abstracts (more coherent, more flexible variable-length abstracts), we argue that the use of human-selected extraction candidates is advantageous to the task. Our favourite heuristic includes meta-linguistic cue phrases, because they can be used to detect rhetorical structure in the document, and because they provide a rhetorical context for each extracted sentence in addition to its propositional content.

6 Acknowledgements

The authors would like to thank Chris Brew, Janet Hitzeman and two anonymous referees for comments on earlier drafts of the paper. The first author is supported by an EPSRC studentship.

References

Baxendale, P. (1958). Man-made index for technical literature – an experiment. *IBM journal on research and development*, 2(4).

Edmundson, H. (1969). New methods in automatic extracting. *Journal of the ACM*, 16(2).

Johnson, F. C., Paice, C. D., Black, W. J., and Neal, A. P. (1993). The application of linguistic processing to automatic abstract generation. *Journal of Document and Text Management*, 1(3):215–42.

Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference.*

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2).

Mann, W. C., Matthiesen, C. M. I. M., and Thompson, S. A. (1992). Rhetorical structure theory and text analysis. In Mann, W. C. and Thompson, S. A., editors, *Discourse description.* J. Benjamins Pub. Co., Amsterdam.

Paice, C. D. and Jones, A. P. (1993). The identification of important concepts in highly structured technical papers. In *Proceedings of the Sixteenth Annual International ACM SIGIR conference on research and development in IR.*

Zechner, K. (1995). Automatic text abstracting by selecting relevant passages. Master's thesis, Centre for Cognitive Science, University of Edinburgh.