A Trainable Document Summarizer

Julian Kupiec, **Jan Pedersen** and **Francine Chen**
Xerox Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304
{kupiec,pedersen,fchen}@parc.xerox.com

Abstract

- To summarize is to reduce in complexity, and hence in length, while retaining some of the essential qualities of the original.
- This paper focusses on document extracts, a particular kind of computed document summary.
- Document extracts consisting of roughly 20% of the original can be as informative as the full text of a document, which suggests that even shorter extracts may be useful indicative summaries.
- The trends in our results are in agreement with those of Edmundson who used a subjectively weighted combination of features as opposed to training the feature weights using a corpus.
- We have developed a trainable summarization program that is grounded in a sound statistical framework.

Keywords: summary sentence, original documents, summary pairs, training corpus, document extracts

1 Introduction

To summarize is to reduce in complexity, and hence in length, while retaining some of the essential qualities of the original. Titles, keywords, tables-of-contents and abstracts might all be considered as forms of summary, however a *document summary* conventionally refers to an abstract-like condensation of a full-text document. Traditionally, document summaries are provided by the author. This paper focusses on *document extracts*, a particular kind of computed document summary.

Abstracts are sometimes used as full document surrogates, for example as the input to text search systems, but they also speed access by providing an easily digested intermediate point between a document's title and its full text that is useful for rapid relevance assessment. It is this second interface-related use that is our motivation for automatic document summarization. The goal is to generate a concise document description that is more revealing than a title but short enough to be absorbed in a single glance. A traditional author-supplied indicative abstract clearly fulfills this objective, but it is hoped that other, more easily computed condensations may also serve.

Numerous researchers have addressed automatic document summarization (see [10] for an overview). The nominal task of generating a coherent narrative summarizing a document is currently considered too problematic since it encompasses discourse understanding, abstraction, and language generation [6]. Nonetheless, knowledge intensive methods have had some success in restricted domains [11, 5, 3, 13, 18]. For example, a filled template produced by a message understanding system can be thought of as a targetted document summary. A simpler, more generic approach avoids the central difficulties of natural language processing by redefining the task to be *summary by extraction* [7]. That is, the goal is to find a subset of the document that is indicative of its contents, typically by scoring sentences and presenting those with the best scores. These sorts of summaries are not guaranteed to have narrative coherence, yet may be useful for rapid relevance assessment.

Document extracts consisting of roughly 20% of the original can be as informative as the full text of a document [9], which suggests that even shorter extracts may be useful indicative summaries. However, other studies [12, 2] suggest that the optimal extract can be far from unique. Numerous heuristics have been proposed to guide the selection of document extracts [7, 4, 17, 14], yet no clear criterion has been proposed to choose among them. Existing evidence [4] suggests that combinations of individual heuristics have the best performance.

We approach extract selection as a statistical classification problem. Given a training set of documents with hand-selected document extracts, develop a classification function that estimates the probability a given sentence is included in an extract. New extracts can then be generated by ranking sentences according to this probability and selecting a user-specified number of the top scoring ones.

This framework provides a natural evaluation criterion: the classification success rate or *precision*. It also offers a direct method for finding an optimal combination of extraction selection heuristics, or features. However, it does require a training corpus of documents with labelled extracts, which can be expensive to obtain. We have acquired such a corpus from Engineering Information Co., a nonprofit company providing abstracts of technical articles to online information services, which will serve as the basis for the experiments described here.

The following sections detail our approach, describe the training corpus, present evaluation results that rate our document summarization method at 42% average precision, and discuss some practical implementation issues.

18th Ann Int'l SIGIR'95/Seattle WA USA-6/95

Aerospace America	Manufacturing Engineering
American Laboratory	Metal Finishing
Civil Engineering	Modern Plastics
Chemical Engineering Education	Oil and Gas Journal
Concrete International	Pulp and Paper International
IEEE Communications Magazine	Robotics World
IEEE Control System	Scripta Metallurgica et Materiala
Journal of Cellular Plastics	Sensors
Journal of Material Science Letters	Water Engineering and Management
Japanese Railway Engineering	Wire Association International '93
Machine Design	

Table 1: Journals in Corpus

2 A Trainable Summarizer

Extracting summarizers typically compute a score for each sentence in a document and then select the highest scoring subset. The scoring criteria employed include participation in predefined semantic roles[11], rhetorical relations[8], inclusion of phrasal index terms[16], document-specific keyword frequencies[7], location heuristics[1], and the assessment of sentence similarity structure[17, 15]. Methods either assume the document exists in isolation, or in the context of a larger collection, which al lows term weights to depend on corpus statistics[14, 15].

The precise formulation of the scoring rule is heuristic and empirical in nature. However, if one were given a training corpus of documents with matched extracts, it would be natural to approach the problem as one of statistical classification. This would provide a principled method for selecting among potential features, or scoring criteria, and for choosing a weighted combination of these to produce an "optimal" scoring scheme — optimal in the sense of doing the best job possible of predicting the extraction selection performed by human judges given the features and the method of combination. To pursue this approach, we need to establish the set of potential features, the classification method, and a training corpus of document/extract pairs.

2.1 Features

Paice [10] groups sentence scoring features into seven categories. *Frequency-keyword* heuristics use the most common content words as indicators of the main themes in the document. Sentences containing these words are scored using functions of their frequency counts [4, 19]. The *title-keyword* heuristic assumes that important sentences contain content words that are present in the title and major headings of a document. *Location* heuristics assume that important sentences lie at the beginning and end of a document, in the first and last sentences of paragraphs [1, 4], and also immediately below section headings. *Indicator phrases* contain words that are likely to accompany indicative or informative summary material (e.g., "This report..."). A related heuristic involves *cue words*. These may include two sets of "bonus" and "stigma" words [4] which are positively and negatively correlated with summary sentences. Example bonus words are "greatest" and "significant". Stigma words are exemplified by "hardly" and "impossible".

Through experimentation we settled on the following feature set, which are all discrete in nature.

Sentence Length Cut-off Feature: Short sentences tend not to be included in summaries (section headings generally count as short sentences). Given a threshold (e.g., 5 words), the feature is true for all sentences longer than the threshold, and false otherwise.

Fixed-Phrase Feature: Sentences containing any of a list of fixed phrases, mostly two words long (e.g., "this letter...", "In conclusion..." etc.), or occurring immediately after a section heading containing a keyword such as "conclusions", "results", "summary", and "discussion" are more likely to be in summaries. This features is true for sentences that contain any of 26 indicator phrases, or that follow section heads that contain specific keywords.

Paragraph Feature: This discrete feature records information for the first ten paragraphs and last five paragraphs in a document. Sentences in a paragraph are distinguished according to whether they are paragraph-initial, paragraph-final (for paragraphs longer than one sentence) and paragraph-medial (in paragraphs greater than two sentences long).

Thematic Word Feature: The most frequent content words are defined as thematic words (ties for words with the same frequency are resolved on the basis of word length). A small number of thematic words is selected and each sentence is scored as a function of frequency. This feature is binary, depending on whether a sentence is present in the set of highest scoring sentences. Experiments were performed in which scaled sentence scores were used as pseudo-probabilities, however this gave inferior performance.

Uppercase Word Feature: Proper names are often important, as is explanatory text for acronyms e.g., *"... by the ASTM (American Society for Testing and Materials)"*. This feature is computed similarly to the previous one, with the constraints that an uppercase thematic word is not sentence-initial and begins with a capital letter. Additionally, it must occur several times and must not be an abbreviated unit of measurement (e.g., F, C, Kg, etc.). Sentences in which such words appear first score twice as much as later occurrences.

2.2 Classifier

For each sentence s we compute the probability it will be included in a summary $\mathcal{S}$ given the k features $F_j; j = 1...k$, which can be expressed using Bayes' rule as follows:

$$P(s \in \mathcal{S}|F_1, F_2, \ldots F_k) = \frac{P(F_1, F_2, \ldots F_k|s \in \mathcal{S})P(s \in \mathcal{S})}{P(F_1, F_2, \ldots F_k)}$$

Assuming statistical independence of the features:

$$P(s \in \mathcal{S}|F_1, F_2, \ldots F_k) = \frac{\prod_{j=1}^{k} P(F_j|s \in \mathcal{S})\ P(s \in \mathcal{S})}{\prod_{j=1}^{k} P(F_j)}$$

$P(s \in \mathcal{S})$ is a constant and $P(F_j|s \in \mathcal{S})$ and $P(F_j)$ can be estimated directly from the training set by counting occurrences. Note that since all the features are discrete, we can formulate this equation in terms of probabilities rather than likelihoods. This yields a simple Bayesian classification function that assigns for each s a score which can be used to select sentences for inclusion in a generated summary.

3 The Corpus

The training corpus provided by Engineering Information employed in our investigation contains documents without author-supplied abstracts. Abstracts were instead created by professional abstractors by reference to the original. There are 188 document/summary pairs, sampled from 21 publications in the scientific/technical domain (see Table 1). These summaries are mainly indicative, and their average length is three sentences. An example is shown in Figure 1.

Documents were received in the form of photocopies which required scanning and optical character recognition (OCR) to extract their text portions. This process introduced spelling errors and occasional omissions of text. The resulting text files were manually checked, and either rejected due to excessive OCR errors or cleaned-up. Errors and omissions still remain in the files after cleanup, however they are unlikely to affect results. Particular care was taken to ensure that the beginnings and ends of documents were correct, as most summary sentences are located at these places. The average number of sentences per document is 86 (a slightly conservative estimate due to the omissions). Each document was "normalized" so that the first line of each file contained the document title. Text describing author, address etc., between the title and the start of the document proper was removed, as was the bibliography. (Techniques for dealing with more typical text are described in Section 6). The corresponding original text for Figure 1 is shown in Figure 2.

The training strategy outlined in Section 2 assumes that we have document/extract pairs. However, we have in fact manual summary sentences that are "inspired" by particular sentences in the original documents. Thus the summarization task we are addressing is to extract the same set of sentences from a document that an expert might use to make summary text, either verbatim or with minor modification, preserving content.

3.1 Sentence Matching

To proceed with training, we need to obtain a correspondence between the manual summary sentences and sentences in the original document. Sentences from the original documents can be matched to those in the manual summaries in several ways. A *direct sentence match* occurs when a manual summary sentence could either be extracted verbatim from the original, or with minor modifications, preserving the content (as exemplified by Figures 1 and 2). When it is obvious that two or more sentences were used from the original to make a summary sentence, a *direct join* occurs. If it is either obvious or suspected that the expert constructed a summary sentence from a general reading (i.e. using no specific sentence from the original) the summary sentence is labelled *unmatchable*. Individual summary sentences may also be labelled *incomplete* in two situations. The first is when some overlap does exist between a summary sentence and one in the original, but the content of the original is not preserved in the summary sentence. The second is when the summary sentence includes a sentence from the original document, but also contains other information that is not covered by a direct join. Joins may themselves be labelled *incomplete* for the same reasons. Examples of these correspondences are shown in the Appendix. The correspondences were produced in two passes. In the first, an automatic alignment program was used to find the best one-to-one sentence match in the original documents for each summary sentence. These were used as a starting point for the manual assignment of correspondences made in the second pass. Table 2 shows the distribution of the correspondences in the training corpus.

Direct Sentence Matches	451	79%
Direct Joins	19	3%
Unmatchable Sentences	50	9%
Incomplete Single Sentences	21	4%
Incomplete Joins	27	5%
Total Manual Summary sents	568	

Table 2: Distribution of Correspondences

The table indicates that 79% of the summary sentences have direct matches. The 19 direct joins consist of a total of 41 different sentences from original documents. For three summary sentences, the best matching "sentences" in the original appeared to be the corresponding document titles. Nine of the manual summary sentences appeared to contain section headings (e.g. in lists). In eight instances a sentence in the original document was split up to make several sentences in the manual summaries.

4 Evaluation

Since we had insufficient data to reserve a separate test corpus we used a cross-validation strategy for evaluation. Documents from a given journal were selected for testing one at a time; all other document/summary pairs were used for training. Results were summed over journals. Unmatchable and incomplete sentences were excluded from both training and testing, yielding a total of 498 unique sentences. We evaluate performance in two ways.

The first evaluation measure is stringent – the fraction of manual summary sentences that were faithfully reproduced by the summarizer program. It is thus limited by the drawbacks of text excerpting and the highest performance attainable is the sum of all direct sentence matches and all direct joins. Referring to Table 2 this is:

$$\frac{451 + 19}{568} = 83\%$$

A sentence produced by the summarizer is defined as correct here if:

1. It has a direct sentence match, and is present in the manual summary.
2. It is in the manual summary as part of a direct join, and all other members of the join have also been produced (thus all the information in the join is preserved).

For each test document, the trained summarizer produced the same number of sentences as were in the corresponding manual summary. Of the 568 sentences, 195 direct sentence matches and 6 direct joins were correctly identified, for a total of 201 correctly identified summary sentences. The summarizer thus replicates 35% of the information in the manual summaries. This assumes that only one "correct" summary exists for a document which is very unlikely to be the case. Indeed, it has been observed that subjects differ greatly when asked to select summary sentences [2]. In particular, Rath et al. [12] found that extracts selected by four different human judges had only 25% overlap, and for a given judge over time only 55% overlap.

The second evaluation measure is the fraction of the 498 matchable sentences that were correctly identified by the summarizer (it is

The work undertaken examines the drawability of steel wire rod with respect to elements that are not intentionally added to steel. Only low carbon steels were selected for experimentation. During wire drawing, failure-inducing tensile forces are greatest at the center of the wire. This accounts for the classic appearance of ductile failure with the center of the wire failing in a ductile manner.

Figure 1: A Manual Summary

Paragraph 2: The work undertaken examines the drawability of steel wire rod with respect to elements that are not intentionally added to steel. The effect of microstructure was not of interest to the investigation. For this reason, only low carbon steels were selected for experimentation.

.....

Paragraph 4: Once nucleated, these microvoids grow and coalesce, until the wire can no longer support the drawing load and a break occurs. During wiredrawing, failure-inducing tensile forces are greatest at the center of the wire. This accounts for the classic appearance of ductile failure with the center of the wire failing in a ductile manner, while the circumference fails last, and in shear.

Figure 2: Relevant Paragraphs from Original

thus theoretically possible to attain 100% correct). When the summarizer outputs the same number of sentences as in corresponding manual summaries, 211 out of 498 (42%) were correctly identified.

The second column in Table 3 shows the sentence-level performance for individual features. In cases where sentences have the same probability, they are ranked in document order. Thus, the sentence length cut-off feature, if used alone, returns the text at the beginning of a document, excluding the title and headings.

Feature	Individual Sents Correct	Cumulative Sents Correct
Paragraph	163 (33%)	163 (33%)
Fixed Phrases	145 (29%)	209 (42%)
Length Cut-off	121 (24%)	217 (44%)
Thematic Word	101 (20%)	209 (42%)
Uppercase Word	100 (20%)	211 (42%)

Table 3: Performance of Features

The third column in Table 3 shows how performance varies as features are successively combined together, in descending order of individual performance. The best combination is (paragraph + fixed-phrase + sentence-length). Addition of the frequency-keyword features (thematic and uppercase word features) results in a slight decrease in overall performance.

For a baseline, we compared the summarizer with the strategy of simply selecting sentences from the beginning of a document (how documents are typically displayed and read). This baseline was computed by considering the sentence length cut-off feature alone, which ranks sentences in reading order, excluding short fragments, such as section headings. When compared to the baseline (which can be read off the third row of Table 3; 121 sentences correct) using the full feature set improves performance by 74% (211 sentences correct).

Figure 3 shows the performance of the summarizer (using all features) as a function of summary size. When generating summaries that automatically select 25% of the sentences in the original documents, Edmundson cites a sentence-level performance of 44%. By analogy, 25% of the average document length (86 sentences) in our corpus is about 20 sentences. Reference to the table indicates performance at 84%.

5 Discussion

The trends in our results are in agreement with those of Edmundson [4] who used a subjectively weighted combination of features as opposed to training the feature weights using a corpus. He also found that location-based heuristics gave best performance. His best combination of heuristics were based on location, title-keywords and cue words. Edmundson also experimented with a frequency-keyword heuristic, omitting it from his preferred selection on account of inferior performance.

Frequency-keyword features (i.e. the thematic word feature and uppercase feature) also gave poorest individual performance in our evaluation. The likely reason is that they select sentences more evenly throughout a text, but our corpus contains a lot of indicative material located at the beginnings and ends. We have however retained these features in our final system for several reasons. The first is robustness; many text genres do not contain any of the indicator-phrases that are common in the corpus we have used [1]. Secondly, as the number of sentences in a summary grows, more dispersed informative material tends to be included.

As described in Section 3.1, we first used an automatic alignment program to obtain correspondences, which were then manually checked and corrected. We also evaluated performance using the manually corrected correspondences, but training using only the correspondences produced by the alignment program. The performance was 216 sentences (43%) correct, suggesting that for corpora such ours, summarizers can be trained automatically from document/summary pairs without manual intervention.

6 Implementation Issues

Our goal is to provide a summarization program that is of general utility. This requires attention to several issues beyond the training of features and performance evaluation. The first concerns robustness (in this regard multiple features have already been discussed).

[1] When the fixed-phrase feature is omitted, performance drops from 211 sentences (42%) to 178 (36%)

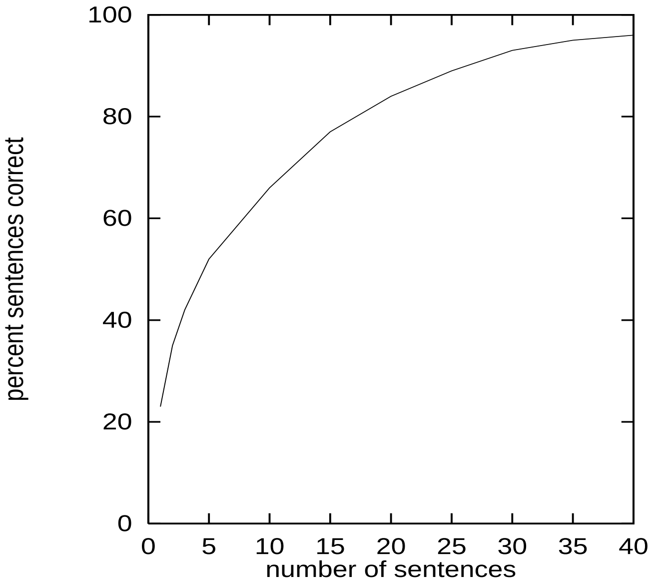

Figure 3: Performance vs. Summary Size

Key Phrases:

cold work	solute atmospheres
solute atoms	test piece
dislocation velocity	dynamic strain aging

Sentence Extracts:

- Drawability of low carbon steel wire
- The work undertaken examines the drawability of steel wire rod with respect to elements that are not intentionally added to steel.
- For this reason, only low carbon steels were selected for experimentation.
- During wiredrawing, failure-inducing tensile forces are greatest at the center of the wire.
- This accounts for the classic appearance of ductile failure with the center of the wire failing in a ductile manner, while the circumference fails last, and in shear.

Figure 4: Computed Summary

As mentioned earlier, documents in the corpus were edited so that the title appears first and the text of the document proper immediately follows. In practice, both the title (if present) and the beginning of the main body of text are often preceded by dates, addresses, names, and various other notations. It is advantageous to find the title, and the beginning of the main text (performance is sensitive to the beginning of the main text, by virtue of the paragraph feature). We therefore implemented another set of features specifically to find the start of the main text body, and to isolate a sentence that acts as a title, lying between the main text and beginning of the document. Briefly, these features include numbers, explicit sentence boundary marks, word case, and paragraph and sentence lengths. For example, uppercase or word-initial uppercase letters are often used in titles, and consecutive sentences ending with explicit punctuation are more likely to be in the main text body. Additionally, if an author-supplied abstract is present (identified by a heading containing the word *abstract*), then subsequent paragraphs are used directly as the summary and no feature-based extraction is attempted.

The second issue concerns presentation and other forms of summary information. The highest scoring sentences (including the likely title) are shown in reading order to the user in conjunction with the key phrases of the document (as illustrated in Figure 4). These key phrases must contain at least two adjacent words, are primarily noun phrases, and are presented in frequency order. They are computed based on a frequency analysis of word sequences in a document. To identify them, a stop list composed of articles, prepositions, common adverbs, and auxilary verbs is used to break the words in a sentence into phrases.

7 Conclusions

We have developed a trainable summarization program that is grounded in a sound statistical framework. For summaries that are 25% of the size of the average test document, it selects 84% of the sentences chosen by professionals. For smaller summary sizes an improvement of 74% was observed over simply presenting the beginning of a document. We have also described how extracts can be used with other information to create a summary useful of rapid relevance assessment while browsing.

8 Acknowledgments

We would like to thank Meg Withgott and John Tukey for their insights into this work. In particular, Meg Withgott was responsible for pioneering this work. We thank Austen Briggs for hand correcting the OCR data. We are also grateful to Dan Brotsky, Ken Feuerman, Sandy Garcia and Steve Putz for their help and for the implementation of the summarization program in the XSoft Visual Recall^{TM} product.

References

[1] P. B. Baxendale. Man-made index for technical literature – an experiment. *IBM J. Res. Develop.*, 2(4):354–361, 1958.

[2] F.R. Chen and M.M. Withgott. The use of emphasis to automatically summarize a spoken discourse. In *Proceedings of the IEEE Intl. Conf. on Acoust., Speech and Signal Proc.*, volume 1, pages 229–232, March 1992.

[3] G. DeJong. An overview of the FRUMP system. In W.G. Lehnert and M. H. Ringle, editors, *Strategies for Natural Language Parsing*, pages 149–176, 1982.

[4] H. P. Edmundson. New methods in automatic abstracting. *Journal of the ACM*, 16(2):264–285, April 1969.

[5] P.S. Jacobs and L. F. Rau. Scisor: Extracting information from on-line news. *Communications of the ACM*, 33(11):88–97, 1990.

[6] K. Sparck Jones. Discourse modelling for automatic summarising. Technical Report 29D, Computer Laboratory, University of Cambridge, 1993.

[7] H.P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Develop.*, 2:159–165, 1959.

[8] S. Miike, E. Itoh, K. Ono, and K. Sumita. A full-text retrieval system with a dynamic abstract generation function. In W. Bruce Croft and C.J. van Rijsbergen, editors, *Proceedings of Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 152–161, July 1994.

[9] A. H. Morris, G. M. Kasper, and D. A. Adams. The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, pages 17–35, March 1992.

[10] C. D. Paice. Constructing literature abstracts by computer: Techniques and prospects. *Information Processing and Management*, 26:171–186, 1990.

[11] C. D. Paice and P. A. Jones. The identification of important concepts in highly structured technical papers. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 69–78. ACM Press, June 1993.

[12] G. J. Rath, A. Resnick, and T. R. Savage. The formation of abstracts by the selection of sentences. *American Documentation*, 12(2):139–143, April 1961.

[13] U. Reimer and U. Hahn. Text condensation as knowledge base abstraction. In *IEEE Conf. on AI Applications*, pages 338–344, 1988.

[14] G. Salton, J. Alan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of SIGIR'93*, pages 49–58, June 1993.

[15] G. Salton, J. Allan, C. Buckley, and A. Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264(3):1421–1426, June 1994.

[16] C. Schwarz. Content based text handling. *Information Processing & Management*, 26(2):219–226, 1990.

[17] E. F. Skorokhod'ko. Adaptive method of automatic abstracting and indexing. In *IFIP Congress, Ljubljana, Yugoslavia 71*, pages 1179–1182. North Holland, 1972.

[18] J. I. Tait. Generating summaries using a script-based language analyzer. In L. Steels and J.A. Campbell, editors, *Progress in Artificial Intelligence*, pages 312–318. Ellis Horwood, 1985.

[19] L. C. Tong and S. L. Tan. A statistical approach to automatic text extraction. *Asian Library Journal*.

9 Appendix

9.0.1 Direct Match

If a summary sentence is identical to a sentence in the original, or has essentially the same content, the match is defined as a direct match. An example match that is not exact but considered to convey the same content is shown below:

Manual: This paper identifies the desirable features of an ideal multisensor gas monitor and lists the different models currently available.

Original: The present part lists the desirable features and the different models of portable, multisensor gas monitors currently available.

9.0.2 Direct Join

If the content of the manual sentence is represented by two or more sentences in the original, the latter sentences are noted as joins. For example:

Manual: In California, Caltrans has a rolling pavement management program, with continuous collection of data with the aim of identifying roads that require more monitoring and repair.

Original (1): Rather than conducting biennial surveys, Caltrans now has a rolling pavement-management program, with data collected continuously.

Original (2): The idea is to pinpoint the roads that may need more or less monitoring and repair.

9.0.3 Incomplete Matches

A sentence in the original document is labelled as an incomplete match if it only partially covers the content of a manual summary sentence, or if a direct match is not clear. It can occur in the context of a single sentence or a join. The following exemplifies an incomplete single sentence match:

Manual: Intergranular fracture of polycrystalline Ni_3Al was studied at 77K.

Original: Before discussing the observed deformation and fracture behavior of polycrystalline Ni_3Al at 77K in terms of the kinetics of the proposed environmental embrittlement mechanism, we should ask whether the low temperature by itself significantly affects the brittleness of Ni_3Al.