An Evaluation of Text Matching Systems for Text Excerpts of Varying Scope

Gerard Salton
Chris Buckley

TR 90-1134
June 1990

Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

*This study was supported in part by the National Science Foundation under Grant IRI-87-02735.

An Evaluation of Text Matching Systems for Text Excerpts of Varying Scope

Gerard Salton and Chris Buckley*

May 31, 1990

Abstract

When large text collections must be processed, it is not possible to limit the scope of the subject matter of interest. In such a situation the standard content analysis methods that are based on the use of knowledge bases to represent the relevant subject areas are not applicable. Necessarily, the texts themselves must then serve as the main basis for the content analysis operations.

Experiments are described in this note designed to evaluate text matching operations for text excerpts of varying scope, including in particular text paragraphs and text sentences extracted from book size materials. The evaluation shows that when the global text similarity between distinct text paragraphs is high, while at the same time local similarities also exist for particular text sentences included in these paragraphs, the presumption is that the paragraphs cover related subject matter. One concludes that text matching systems may prove useful for text linking and information retrieval.

1 Introduction

In many text processing environments, it is necessary to deal with large collections of natural language texts – for example, collections of textbooks, or volumes of machine-readable encyclopedias and dictionaries. In such circumstances, it is not possible to limit the scope of the subject matter of interest, and the standard content analysis methods that are based on the use of thesauruses and knowledge bases to represent the main concepts and the principal

*Department of Computer Science, Cornell University, Ithaca, NY 14853-7501. This study was supported in part by the National Science Foundation under grant IRI 87-02735.

term relationships do not apply. Indeed not enough is known about the generation of knowledge bases of unlimited scope to make it sensible to pursue an approach using such tools.

While it is difficult to build artificial constructs for the representation of document content, large samples of machine-readable text can be processed directly on a computer in a straightforward way. This suggest that a content analysis system might be based directly on a study of the original texts themselves, rather than on the construction of artificial tools such as dictionaries or thesauruses. In an earlier report, global vocabulary matching approaches were described that recognize homogeneous text environments, based on coincidences between the vocabularies used in different contexts.[1] In that study, the usual text analysis operations which resolve ambiguities in text meaning were replaced by multiple text comparisons between excerpts of varying scope. It was found that coincidences in text meaning could be detected in a large majority of cases, when local as well as global similarities exist between particular text excerpts.

In the earlier study, the full text of a standard textbook was analyzed, and paragraph as well as sentence similarities were computed for approximately 650,000 paragraph pairs and over 10 million sentence pairs.[2] Among the 1474 paragraph pairs that exhibited both high global (paragraph) similarities as well as high local (sentence) similarities, only 56 pairs appeared to be improperly related. That is, for 1418 out of 1474 matching paragraph pairs (96 percent of the paragraph pairs), the subject matter of the respective pairs was found to be identical.[1]

In the present note, various questions are considered that were left unexplained earlier. The following problems are of particular interest:

a) The methods used to represent and compare the sentence and paragraph pairs.

b) The usefulness of restricting sentence comparisons to sentences located in particular paragraph pairs – for example, paragraphs exhibiting substantial global text similarities.

c) The importance of granularity in the text matching system; in particular, the advantages of using either global paragraph similarities or local sentence similarities in text retrieval.

d) The effectiveness in retrieval of local sentence similarities used in addition to global paragraph similarities.

The evaluation data shown in this note are based in each case on a detailed examination of certain document pairs, extracted from a list of all document pairs arranged in decreasing order of pairwise text similarity. In particular, all available text excerpts (text paragraphs or text sentences) are first compared with each other, and a similarity coefficient is obtained for each text pair. The

text pair output is then arranged in decreasing order of pairwise similarity, and the texts of certain document pairs are manually examined to determine whether sufficient similarities exist between the respective subject areas. Search precision figures are then computed reflecting the proportion of text pairs that were correctly linked by the automatic vocabulary matching system. If the automatic text comparison system operates as expected, one would expect that many of the document pairs with high similarity coefficients (that is, with low output ranks) do in fact cover identical subject matter, and that the subject similarity decreases as the text similarity coefficients become smaller.

In the evaluation output that follows, 60 text pairs are examined for each collection, including 20 text pairs retrieved early in the search (that is, with high pairwise similarity), 20 pairs retrieved with retrieval ranks of about 550, and 20 pairs with ranks of about 1,100. For the better text comparison methods, a large proportion of the 60 examined text pairs was found to be appropriately linked. Detailed evaluation results are shown in the remainder of this note.

2 Automatic Text Matching

A) General Considerations

Consider first the problem of text representation for text matching purposes. The standard text indexing systems that have proven themselves in the past use the following main steps:[2]

- identification of individual text words;
- elimination of common words included on a special list of excluded words;
- reduction to word stem form by suffix removal;
- weighting of word stems;
- global comparison of sets of weighted word stems.

The first three steps are normally considered noncontroversial. The elimination of the function words reduces the text size by about 50 percent, and the reduction to word stem form further decreases the size of the indexing vocabulary.

The term weighting step does, however, raise important problems that may crucially affect the text matching system. Consider, as an example, the two sample sentences of Fig. 1, extracted from the text of reference [2]. The matching terms for the two texts are underlined. If a raw (term frequency) count of matching terms is used to measure the similarity between the sample texts, a total similarity of 9 is obtained, produced by the coincidence between the following nine word stems:

attach, attrib, exact, match, query, record, retrief, stor, valu.

Special provisions may be made for adding phrase identifiers to the single-term identifiers. If a phrase is defined as a pair of matching words occurring adjacently in the document text, three phrases are detected for the sample texts, that is "attrib valu", "exact match", and "stor record". If each phrase match is counted as an additional term match, a final text similarity coefficient of 12 is obtained for the sample texts of Fig. 1.

When a raw term (word stem) match is used to determine text similarity based on the number of matching terms or phrases, higher text similarity coefficients are normally obtained for longer documents with more included terms than for shorter texts. Indeed, when more terms are initially available, more matching terms are normally detected. In many text environments, the document length may vary considerably, even though intrinsically each text will be equally valuable for retrieval purposes. A rule favoring the retrieval of longer documents over shorter ones might then produce disastrous retrieval results. Such a situation can be remedied by assigning normalized term weights to the document identifiers, chosen in such a way that the total document length is the same for all documents. A typical term weight of this kind is

$$w_{ij} = \frac{tf_{ij}}{\sqrt{\sum_j (tf_{ij})^2}}$$

where w_{ij} represents the weight of term j in document D_i and tf_{ij} is the frequency of occurrence of term j in D_i.

When normalized term weights are used, and the similarity between two texts D_i and D_k is computed as the sum of the products of coinciding terms (similarity $(D_i, D_k) = \sum_j w_{ij}.w_{kj}$), the pairwise similarity between texts ranges between 0 and 1, reflecting the situations when no common terms exist (similarity = 0), and when all terms occur in both documents with equal weight (similarity = 1). When some of the terms coincide but others do not, the pairwise similarity will be greater than 0 but less than 1.

A collection-dependent frequency factor, known as the inverse document frequency (idf), may be included as part of the term weight, reflecting the fact that terms useful for content identification occur frequently in particular local documents, but rarely on the outside. A standard composite term weight is then computed by multiplying the term frequency and inverse document frequency factors ($tf \times idf$), and normalizing for document length. [1,2]

B) Paragraph and Sentence Comparisons

The effectiveness of the text matching operations is evaluated by comparing the text paragraphs and sentences, and computing search precision figures reflecting the proportion of properly related texts that were jointly identified by the text matching system. In the present study, the full text of reference [2] is used for experimental purposes. Several different indexing and text matching methods are used to process the paragraphs and sentences, and output evalua-

tion figures are computed to assess system effectiveness.

A list of the document collections used experimentally appears in Table 1. Collections 21 and 24 cover the 2,400 paragraph pairs with the highest pairwise text similarities, out of nearly 650,000 distinct paragraph pairs that are defined for the original 1,137 paragraphs of reference [2]. A normalized ($tf \times idf$) term weighting system is used to construct collection 21 with normalized document lengths. On the other hand, standard term frequency weights serve for collection 24 where the text similarities are based on the number of matching term pairs. When term frequency weights control the text similarity measurements, the longer texts that contain more terms, and terms with higher weights, will receive higher similarity coefficients than shorter texts with fewer terms. In a retrieval situation where text paragraphs, and other longer text excerpts, are processed and all paragraphs are assumed to be equally valuable, normalized text similarity measures may prove to be more useful than measurements that depend on the document length and favor the longer text excerpts.

Five collections of sentence pairs are also shown in Table 1, labelled 3, 17, 19, 31, and 35, respectively. For collections 3, 17, and 19, the 2,400 sentence pairs are chosen from the complete set of over 10 million sentence pairs that can be generated for the 4,500 text sentences of reference [2]. On the other hand, the sentences used in collections 31 and 35 are all extracted from pairs of paragraph known to be included in paragraph collection 21. That is, for each sentence pair included in collections 31 or 35, it is known that a corresponding matching paragraph pair exists in collection 21.

Term frequency weights are used for all sentence collections, except collection 19, and sentence similarities are computed by using the number of matching term pairs rather than normalized term frequency measures. The reason is that the sample texts contain many short sentences, consisting for example of section or chapter titles. Short sentences are identified by very few terms – possibly only one or two – and large numbers of such sentence pairs exhibit perfect similarities of 1 in a normalized sentence matching system. For example, in sentence collection 19, consisting of the best 2,400 sentence pairs using the indexing method of collection 3 with normalized ($tf \times idf$) sentence similarity measures, 303 sentence pairs have perfect pairwise similarity of 1. More often than not, these sentence matches are not significant for content description. Table 2 contains a variety of sentence pairs consisting mostly of fragments such as section titles, illustrations, and introductions to figures. In each case, large sentence similarities are obtained when normalized term weight measurements are used, even though the sentence similarity often depends on only one or two nonsignificant words, such as "figure" or "example".

For short text excerpts, such as phrases or sentences, it is therefore reasonable to use a text matching method which favors the longer sentences by requiring a minimum number of matching terms before designating the sentences as related. This is achieved by using term matches with term frequency weights as a matching criterion.

A number of different text indexing systems are used with the collections of Table 1. For collection 3, 19, 21 and 24, all significant word stems included in the text (after deletion of common terms) are used for indexing purposes, and phrases are not formed. Collection 17 uses only the 10 best terms (in decreasing term weight order) from each document to index the corresponding text sentences; phrases formed from adjacent significant terms are used in addition to single terms. Finally, for sentence collections 31 and 35, the matching terms from the document pair containing the sentences are used for indexing; additionally, term phrases are formed for collection 31 but not for 35.

It should be noted that the text matching operations appear not to be grossly affected by minor differences in indexing policies. Table 3 shows that there is a perfect overlap for paragraph collections 21 and 25, and for sentence collections 31 and 35, for document pairs retrieved early in a search (output ranks 1-21). The overlap is still very substantial for texts retrieved with ranks of about 500, and about 1,000. This explains why nearly identical text linking results are obtained for collection pairs that use somewhat different indexing policies but identical text similarity measurements.

3 Evaluation Results

A) Pairwise Paragraph and Sentence Linking

It was suggested earlier that normalized text similarity measurements may be preferred for larger text excerpts, such as paragraphs, because each text is then treated as equally valuable. On the other hand, when short text fragments must be processed, a minimum number of matching terms may be needed for retrieval to insure that nonsignificant text fragments are properly rejected. This is confirmed by the evaluation output contained in Tables 4 and 5 for global paragraph matching and sentence matching systems, respectively.

Table 4 contains an evaluation of the accuracy of text linking for paragraph collections 21 and 24. The assessment is based on a detailed examination of 180 document pairs, including 60 exhibiting the highest pairwise similarity coefficients when the documents are arranged in decreasing order of the pairwise similarity, 60 with a rank around 550 in the ranked order, and 60 with a rank around 1,100. For each set of 60 paragraph pairs, 20 pairs chosen from the output of collection 21 are not also contained in collection 24 (that is, these paragraph pairs appear among the best 2,400 pairs for collection 21 but not for collection 24); another 20 are taken from the output of collection 24 that are not also included in collection 21. The last 20 appear in both collections.

Table 4 shows that the text similarity measure used for collection 21 with length normalization is far superior to that used with collection 24. For collection 21, 53 paragraph pairs out of the 60 that were examined cover reasonably similar texts (acceptable pairs). In addition, the meanings were completely identical (correct pairs) for 46 pairs out of the 53 acceptable pairs. On the other

hand, in collection 24, only 20 pairs out of the 60 examined ones were properly linked. A comparison of the two right-most columns of Table 4 also indicates that the additional use of the term frequency weighting system available with collection 24 does not improve the linking accuracy obtainable with the normalized ($tf \times idf$) system of collection 21. In fact, the results obtained for collection 21, which reach a linking accuracy of 100% for paragraph pairs identified early in a search, and an overall accuracy of 88%, is not improved by using document pair that are jointly retrieved in both collections.

In interpreting the precision results of Table 4, it must be remembered that 40 out of the 60 evaluated pairs appear far down on the output list with retrieval ranks of about 550 for twenty pairs and about 1100 for the remaining twenty pairs. Given that the retrieval accuracy in collection 21 is still around 80 percent for documents pairs with low similarity coefficients and ranks around 1,100, one concludes that the corresponding text matching system provides a high order of performance.

When text sentences are compared instead of paragraph-length text excerpts, the text similarity is best computed by using the number of matching term pairs as a retrieval criterion. Table 5 contains a comparison of term weighting and text similarity measurements for sentence (rather than paragraph) matches. Sentence collections 3 and 19 are used for the output of Table 5. It may be seen that the term frequency weights used with collection 3 give much better sentence matching results than the normalized weighting schemes used for collection 19. Overall, only about 17 percent of the sentence links are acceptable for collection 19, compared with 57 percent of the links in collection 3. By far the best sentence matching results are obtained by using a combination of techniques requiring both a high global (normalized) similarity, as well as a reasonable number of individual term matches. The right-most column of Table 5 shows that 85 percent of the sentence links are acceptable for sentence pairs that are linked by both the strategies used for collections 3 and 19.

In a sentence matching environment, it is also useful to compare the overall accuracy of a global comparison of sentence pairs with the accuracy of sentence comparisons for sentences known to occur in already matching paragraph pairs. The corresponding evaluation data are shown in Table 6, where collection 17 covers the global sentence pair match using the top 2,400 sentence pairs (out of over ten million pairs) regardless of sentence origin, and collection 31 includes only sentences from document pairs in collection 21. The output of Table 6 shows that the overall accuracy of sentence matching reaches only about 30 percent for the global comparisons of collection 17, whereas a reasonably high accuracy of 82 percent is obtained for collection 31. Indeed, the accuracy of sentence pair linking is 100 percent in collection 31 for documents retrieved early in the search. Once again the joint use of the global sentence comparisons as well as comparisons within already matching documents (joint output for collections 17 and 31) does not produce improvements over the use of collection 31 alone.

In summary, it appears that reasonably high linking accuracies are obtainable for global paragraph comparisons using ($tf \times idf$) term weighting strategies and normalized text similarity measurements. A somewhat lower but still respectable retrieval accuracy is obtainable for sentence comparisons based on the number of matching terms pairs when the sentence pairs are chosen from within already matching paragraph pairs.

B) Paragraph Retrieval Using Global as Well as Local Text Matching

In environments where local text excerpts can be compared, such as for example text phrases and sentences, retrieval strategies can be considered that retrieve higher-order constructs, such as text paragraphs, when a sufficient degree of coincidence exists between included lower-level structures. For example, a high degree of similarity between the sample sentences 467-01 and 772-00 of Fig. 1 (sentences 01 of document 467 and 00 of document 772) may lead to the joint retrieval of the corresponding document pair (467 and 772).

Two main questions are investigated in this connection:

a) Should paragraph linking and retrieval be based on global comparisons of complete paragraphs or on local comparisons of included lower-level constructs?

b) Are paragraph comparisons more secure when the corresponding paragraphs also contain matching lower-level constructs such as matching text sentences?

The first question is examined by considering the output of Tables 7 and 8, covering comparisons of collections 3sd and 21, and 3sd and 24, respectively. Collection 3 sd covers paragraph pairs derived from the matching sentence pairs included in collection 3 (as illustrated in the earlier example for sentences 467-01 and 772-00). The output of Table 7 shows that the global paragraph comparisons provide more reliable text linking and retrieval accuracies than the corresponding sentence comparison methods. For highly matching text pairs that are retrieved early in a search operation, the paragraph linking system produces very high accuracy – a fact noted already in connection with the output of Table 4. The corresponding sentence retrieval system for highly matching sentences provides search precision levels of only about 65%. Overall the accuracy of paragraph comparisons is 77% for global text comparisons, but only 58% for the sentence comparisons that form the basis for paragraph retrieval. When global sentence comparisons are used in addition to global paragraph comparisons, the overall retrieval accuracy reaches 85%. This indicates that the text linking accuracy is improved by carrying out text linking operations using several different text environments, and merging the corresponding output results.

This conclusion is reinforced by considering the output of Table 8 where a relatively inferior global paragraph comparison (collection 24) is used, in addition

to the global sentence comparison of collection 3 sd. Here the overall retrieval accuracy provided by both sentence comparisons and paragraph comparisons is approximately the same, but the joint effectiveness for both methods reaches impressive levels of 90 percent. One concludes that local sentence matches can be used to improve the precision of paragraph linking, especially when a weaker system of paragraph matching is used.

In Tables 7 and 8, global sentence matches are used to retrieve the corresponding paragraph pairs. The sentence comparisons are not, however, restricted to particular sentence pairs specified in advance. When the aim is paragraph retrieval or paragraph linking, it appears sensible to restrict the sentence comparisons to sentence pairs known to occur in the relevant pairs of the corresponding paragraph collection. Such a restriction is inherent in the use of collection 31 sd to replace 3 sd. The corresponding evaluation results are given in Table 9.

Table 9 shows that the joint output for paragraphs also containing matching sentences (21 and 31 sd) is far superior to the output obtained with the global paragraph matches alone (21 only). Even for document pairs with ranks above 1000, the text linking accuracy reaches 85 percent when matching sentences are present. For lower ranked paragraph pairs retrieved earlier in the search, the linking accuracy is nearly perfect. The second column of Table 9 is empty because all document pairs specified in collection 31 sd are known in advance to occur also in collection 21.

It was noted earlier that small differences in text indexing will not substantially affect the text linking operations. This is confirmed by comparing the output of Tables 9 and 10. In Table 10 collection 35 sd is used to retrieve paragraph pairs with matching text sentences, instead of collection 31 sd used in Table 9. In the latter case, term phrases are added to the single terms to represent sentence content, whereas in collection 35, the sentences are indexed using the single terms without phrases. The data of Tables 9 and 10 show that the retrieval accuracy is nearly identical in both cases. Once again, a very high order of retrieval precision is obtained for text pairs that exhibit high global (paragraph) similarities, as well as high local (sentence) similarities.

A final analysis for paragraph pairs containing matching text sentences appears in Table 11, covering comparisons between the global paragraph pairs of collection 24, and the paragraph pairs containing matching sentences of collection 31 sd. In this case, different text indexing and retrieval methods are used for the two collections: collection 31 sd is a subset of collection 21 based on $(tf \times idf)$ term assignments, whereas collection 24 uses tf weights alone. The right-hand column of Table 11 shows that the text linking results are practically perfect when the strengths of different systems are superimposed. The output listed in the right-hand column of Table 11 represents paragraph pairs that have both a high global $(tf \times idf)$ normalized text similarity (collection 31 sd), as well as the large number of matching term pairs that is inherent in the similarity measure used with collection 24. These document pairs also contain at least one

pair of matching text sentences.

One hopes that the high order of text matching accuracy available with the sample collection of textbook paragraphs can be obtained also for more heterogeneous text collections than those used in the present experiments.

References

1. G. Salton and C. Buckley, Approaches to Global Text Analysis, Technical Report TR. 90-1113, Computer Science Department, Cornell University, Ithaca, NY, April 1990.

2. G. Salton, Automatic Text Processing, Addison-Wesley Publishing Company, Reading MA, 1989.

Step 1: Sentence 467-01

The retrieval of records depends on an exact match between the attribute values used in the query formulations and those attached to the records being sought: each retrieved record will contain the precise attribute values specified in the query... while each nonretrieved record exhibits at least one mismatch between attribute values attached to the query and those attached to the stored records.

Step 2: Sentence 772-00

Retrieval strategies for various document components differ because usually an exact match is required between the data portions contained in the query specifications and the attribute values attached to the stored records...

Figure 1: Two Sample Sentences from Text of Reference [2]

Collection Number	Sentence or Document Pairs	Type of Indexing	Type of Similarity Measure
21	2400 document (paragraph) pairs	all word stems; no phrases	Normalized ($tf \times idf$) comparison
24	2400 document (paragraph) pairs	all word stems; no phrases	Term frequency (tf); comparison by number of matching terms
3	2394 sentence pairs	all word stems; no phrases	Term frequency (tf) weights; number of matching terms ≥ 7
17	2400 sentence pairs	10 best weighted terms from document plus phrases	Term frequency (tf) weight; comparison by number of matching terms
19	2400 sentence pairs	all word stems; no phrases	Normalized ($tf \times idf$) comparison
31	2177 sentence pairs within top 2400 document pairs of collection 21	all matching terms from document pair plus phrases	Term frequency (tf) weights; number of matching terms ≥ 7
35	2179 sentence pairs within top 2400 document pairs of collection 21	all matching terms from document pair; no phrases	Term frequency (tf) weights; number of of matching terms ≥ 6

Table 1: Experimental Sentence-Pair and Document-Pair Collections

	Sentence Pair		Sentence Similarity (normalized $tf \times idf$ term weights)
a)	828-03:	For the example of Fig.	1.0000
	1017-02:	See for example Fig.	
b)	176-00:	Office Information Retrieval (title)	
	487-04:	information: $R_{345}, R_{348}, R_{350}$ retrieval: $R_{123}, R_{128}, R_{345}$	0.7191
c)	420-01:	There are three possibilities	0.7197
	701-05:	Several possibilities present themselves	
d)	491-00:	Term Weights (title)	
	743-02:	Exclusive dependence on the term of maximum or minimum weight is remedied by introducing term weights for query terms as well as document terms	0.7170
e)	149-04:	Alternatively one can write:	0.7167
	826-05:	11.3(a) would be written as:	

Table 2: Sentence Fragments with Large Pairwise Similarity (examples for collection 19)

Paragraph Collection 21	2400 pairs	all terms no phrases	normalized ($tf \times idf$) comparison
Paragraph Collection 25	2400 pairs	10 best terms plus phrases	normalized ($tf \times idf$) comparison
Paragraph Pairs for Collection 25		Proportion of Paragraph Pairs for Collection 25 also included in Collection 21	
Ranks 1 - 21		21/21	(100%)
Ranks 501 - 521		19/21	(90%)
Ranks 1001 - 1021		16/21	(76%)
Sentence Collection 31	2177 pairs	all matching terms plus phrases	tf weight, number of matching terms
Sentence Collection 35	2179 pairs	all matching terms no phrases	tf weights number of matching terms
Sentence Pairs for Collection 31		Proportion of Sentence Pairs for Collection 31 also included in Collection 35	
Ranks 1 - 21		21/21	(100%)
Ranks 501 - 521		17/21	(81%)
Ranks 1001 - 1021		14/21	(67%)

Table 3: Overlap in Related Text Collections

	Paragraph Pairs retrieved only for collection 24	Paragraph Pairs retrieved only for collection 21	Paragraph Pairs retrieved in common for 21 and 24
Early Output (best ranks)			
correct pairs	10/20 (50%)	18/20 (90%)	18/20 (90%)
acceptable pairs	12/20 (60%)	20/20 (100%)	19/20 (95%)
Medium Output (ranks $\approx$ 550)			
correct pairs	4/20 (20%)	14/20 (70%)	14/20 (70%)
acceptable pairs	6/20 (30%)	17/20 (85%)	17/20 (85%)
Late Output (ranks $\approx$ 1100)			
correct pairs	1/20 (5%)	14/20 (70%)	13/20 (65%)
acceptable pairs	2/20 (10%)	16/20 (80%)	17/20 (85%)
Total			
correct pairs	15/60 (25%)	46/60 (77%)	45/60 (75%)
acceptable pairs	20/60 (33%)	53/60 (88%)	53/60 (88%)

Table 4: Evaluation for Global Paragraph Matching Using Different Text Similarity Measures (Collections 21 and 24)

Proportion of Sentence Pairs with acceptable content links	Sentence Pairs retrieved only for collection 3	Sentence Pairs retrieved only for collection 19	Sentence Pairs retrieved in common for 7 and 19
Early Output (best ranks)	12/20 (60%)	2/20 (10%)	20/20 (100%)
Medium Output (ranks $\approx$ 550)	14/20 (70%)	8/20 (40%)	17/20 (85%)
Late Output (ranks $\approx$ 1100)	8/20 (40%)	0/20 (0%)	14/20 (70%)
Total	34/60 (57%)	10/60 (17%)	51/60 (85%)

Table 5: Evaluation for Global Sentence Matching Using Different Text Similarity Measures (Collections 3 and 19)

	Sentence Pairs retrieved only for collection 31	Sentence Pairs retrieved only for collection 17	Sentence Pairs retrieved in common collections 31 and 17
Early Output (best ranks)			
correct pairs	13/20 (65%)	4/20 (20%)	20/20 (100%)
acceptable pairs	20/20 (100%)	5/20 (25%)	20/20 (100%)
Medium Output (ranks $\approx$ 550)			
correct pairs	19/20 (95%)	1/20 (5%)	14/20 (70%)
acceptable pairs	20/20 (100%)	3/20 (15%)	16/20 (80%)
Late Output (ranks $\approx$ 1100)			
correct pairs	9/20 (45%)	8/20 (40%)	6/20 (30%)
acceptable pairs	13/20 (65%)	10/20 (50%)	13/20 (65%)
Total			
correct pairs	41/60 (68%)	13/60 (22%)	40/60 (67%)
acceptable pairs	53/60 (88%)	18/60 (30%)	49/60 (82%)

Table 6: Evaluation for Global Sentence Comparisons (17-31)

	Paragraph Pairs retrieved only for collection 3sd	Paragraph Pairs retrieved only for collection 21	Paragraph Pairs jointly retrieved for collections 21 and 3sd
Early Output (best ranks)			
correct pairs	9/20 (45%)	19/20 (95%)	19/20 (95%)
acceptable pairs	13/20 (65%)	20/20 (100%)	20/20 (100%)
Medium Output (ranks $\approx$ 550)			
correct pairs	7/20 (35%)	10/20 (50%)	9/20 (45%)
acceptable pairs	10/20 (50%)	12/20 (60%)	15/20 (75%)
Late Output (ranks $\approx$ 1100)			
correct pairs	9/20 (45%)	11/20 (55%)	11/20 (55%)
acceptable pairs	12/20 (60%)	14/20 (70%)	16/20 (80%)
Total			
correct pairs	25/60 (42%)	40/60 (67%)	39/60 (65%)
acceptable pairs	35/60 (58%)	46/60 (77%)	51/60 (85%)

Table 7: Paragraph Retrieval via Global Sentence Comparison (3-21)

	Paragraph Pairs retrieved only for collection 3sd	Paragraph Pairs retrieved only for collection 24	Paragraph Pairs jointly retrieved for collections 24 and 3sd
Early Output (best ranks)			
correct pairs	10/20 (50%)	17/20 (85%)	19/20 (95%)
acceptable pairs	12/20 (60%)	19/20 (95%)	20/20 (100%)
Medium Output (ranks $\approx$ 550)			
correct pairs	9/20 (45%)	7/20 (35%)	14/20 (70%)
acceptable pairs	14/20 (70%)	8/20 (40%)	16/20 (80%)
Late Output (ranks $\approx$ 1100)			
correct pairs	10/20 (50%)	3/20 (15%)	18/20 (90%)
acceptable pairs	14/20 (70%)	8/20 (40%)	18/20 (90%)
Total			
correct pairs	29/60 (48%)	29/60 (48%)	51/60 (85%)
acceptable pairs	40/60 (67%)	35/60 (58%)	54/60 (90%)

Table 8: Paragraph Retrieval via Global Sentence Comparison (3-24)

	Paragraph Pairs retrieved only for collection 31sd	Paragraph Pairs retrieved only for collection 21	Paragraph Pairs jointly retrieved for collections 21 and 31sd
Early Output (best ranks)			
correct pairs	—	17/20 (85%)	19/20 (95%)
acceptable pairs		19/20 (95%)	20/20 (100%)
Medium Output (ranks $\approx$ 550)			
correct pairs	—	10/20 (50%)	14/20 (70%)
acceptable pairs		13/20 (65%)	16/20 (80%)
Late Output (ranks $\approx$ 1100)			
correct pairs	—	9/20 (45%)	17/20 (85%)
acceptable pairs		13/20 (65%)	17/20 (85%)
Total			
correct pairs	—	36/60 (60%)	50/60 (83%)
acceptable pairs		45/60 (75%)	53/60 (88%)

Table 9: Retrieval of Paragraphs Containing Matching Sentences (21-31)