Using Cohesion and Coherence Models for Text Summarization

Inderjeet Mani, Eric Bloedorn, and Barbara Gates
The MITRE Corporation,
11493 Sunset Hills Road, W640, Reston, VA 22090, USA
{imani,bloedorn,blgates}@mitre.org

Abstract

In this paper we investigate two classes of techniques to determine what is salient in a text, as a means of deciding whether that information should be included in a summary. We introduce three methods based on text cohesion, which models text in terms of relations between words or referring expressions, to help determine how tightly connected the text is. We also describe a method based on text coherence, which models text in terms of macro-level relations between clauses or sentences to help determine the overall argumentative structure of the text. The paper compares salience scores produced by the cohesion and coherence methods and compares them with human judgments. The results show that while the coherence method beats the cohesion methods in accuracy of determining clause salience, the best cohesion method can reach 76% of the accuracy levels of the coherence method in determining salience. Further, two of the cohesion methods each yield significant positive correlations with the human salience judgments. We also compare the types of discourse-related text structure discovered by cohesion and coherence methods.

Introduction

As the flood of on-line text information continues unabated, triggered in part by the growth of the World Wide Web, it is especially useful to have tools which can help users digest information content. Text summarization attempts to address this problem by taking a partially-structured source text, extracting information content from it, and presenting the most important content to the user in a manner sensitive to the user's needs. In determining what is salient, the field of text summarization has explored a variety of methods. In this paper, we will focus on two classes of techniques to determine what is salient, based respectively on a representation of text structure in terms of text cohesion and text coherence (Halliday and Hasan 1996).

Text cohesion involves relations between words or referring expressions, which determine how tightly connected the text is. These cohesive relations include repetition, synonymy, anaphora, and ellipsis, and with the exception of ellipsis have been the focus of renewed interest in text summarization (Mani and Bloedorn 1997a), (Mani and Bloedorn 1997b), (Barzilay and Elhadad 1997), (Morris and Hirst 1991), (Boguraev and Kennedy 1997), (Hearst 1994). Models based on coherence (Mann and Thompson 1988), (Liddy 1991), (Van Dijk 1988), on the other hand, represent the overall structure of a multi-sentence text in terms of macro-level relations between clauses (though in some accounts, the elementary text units may be smaller than a clause) or sentences. For example, the connective phrase "in order to", one could argue, expresses some sort of purpose relation between clauses. These relations determine the overall argumentative structure of the text, which are responsible for making the text "cohere". While coherence has always been of interest in synthesizing text from underlying representations, as in text planning for natural language generation (e.g., (Mann and Thompson 1988)), including the generation of summaries, it has recently been the subject of renewed interest (Miike et al. 1994), (Marcu 1997a) as a method of *analyzing* text for text summarization. However, thus far little attention has been paid to systematic studies of methods for integrating coherence and cohesion in text summarization. This paper attempts to address this problem.

Automatic text summarization can be characterized as involving three phases of processing: analysis, refinement, and synthesis. The analysis phase builds a representation of the source text. The refinement phase transforms this representation into a summary representation, condensing text content by selecting salient information. The synthesis phase takes the summary representation and renders it in natural language using appropriate presentation techniques. Models of cohesion and coherence can be considered useful in all three phases.

Conceptually, both coherence and cohesion are somewhat independent of each other. Coherence reflects the deliberate organization of the text by the author in terms of a hierarchical structure to achieve particular argumentative goals. Thus, the nature of the argumentation helps to distinguish a coherent, wellorganized text from one which rambles. Cohesion, on the other hand, is brought about by linguistic devices used to relate different portions of the text, which lend to the text its sense of connectedness. However, there is certainly an overlap between the two, since the sentences or clauses being related in terms of coherence relations describe entities which are themselves related across the text by relations of cohesion.

In order to study this relationship further, one could examine to what extent cohesion and coherence can each be used to establish salience of certain units of text. Then, one could see to what extent the salience predictions of the two types of analyses are correlated with human judgments of salience. In general, this would suggest where cohesion and coherence might work separately or together in determining salience. Having done this, one might go on to examine how the determination of hierarchical structure by coherence can be helped by the use of cohesion relations. This is in fact the approach we will follow.

In what follows, we will first describe our methods of determining salience based on cohesion and coherence, and then go on to describe our experiments comparing the salience predictions of the two.

Cohesion

In general, the relations grouped under 'text cohesion' as used by (Halliday and Hasan 1996) include linguistic devices such as anaphora, ellipsis, conjunction and lexical relations such as reiteration, synonymy, hypernymy, and conjunction. In addition to being a renewed focus of interest among researchers in the text summarization field, various cohesion relations are also of considerable interest in other applications of modeling text discourse structure, including identifying topic shifts (e.g., (Barzilay and Elhadad 1997), (Hearst 1994)) or clustering text (e.g., (Yaari 1997), (Green 1997)).

The cohesion relations we will confine ourselves to here include proper name anaphora, reiteration, synonymy, and hypernymy, as these are relatively easy to define and compute.

Representing Cohesion relations

In order to build a general model of cohesion, it is useful to explicitly represent cohesive relations in text. To that end, we represent a text as a graph, whose nodes represent word instances at different positions. This is in keeping with the assumption underlying much summarization research that location and linear order of text items are important (e.g., (Edmundson 1969), (Kupiec, Pedersen, and Chen 1995), (Paice and Jones 1993)) in determining what's salient. It is also motivated by our goal of generating summary extracts: having a salience function which can tell us which positions are more salient is obviously of interest.

In general, this representation also allows for the representation of different levels of hierarchical structure needed for linguistic analysis, as words in the graph can be grouped into phrases, which can in turn be grouped into clauses and sentences. Obviously, such a representation can represent coherence relations as well, as the sentences and clauses can in turn be grouped into rhetorical structures.

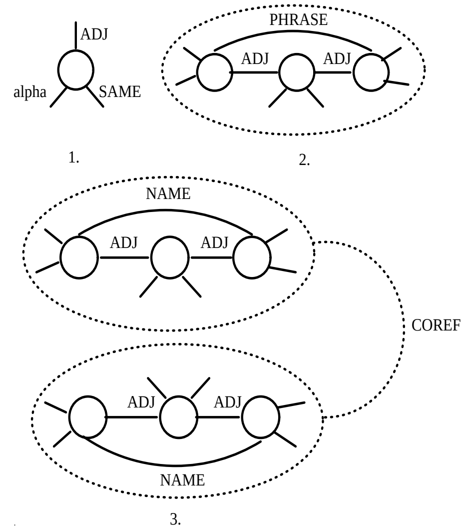

Figure 1: Graph Representation

Given the nodes in the graph, the cohesion relations we extract from the text include repetition, adjacency, synonymy, hypernymy, and coreference among term instances. As shown in Figure 1, each node is a word *instance*, and has a distinct input position. Associated with each such node is a record characterizing the various features of the word in that position (e.g., absolute word position, position in sentence, tf.idf value, part-of-speech). As shown in part 1 of the figure, a node can have adjacency links (ADJ) to textually adjacent nodes, SAME links to other instances of the same word, and other semantic links (represented by *alpha*). PHRASE links tie together strings of adjacent nodes which belong to a phrase (part 2). In part 3, we show a NAME link, as well as the COREF link between subgraphs, relating positions of name instances which are coreferential. NAME links can be specialized to different types, e.g., person, province, etc. In our work, the *alpha* links are restricted to synonymy and hypernymy among words.

The tools used to build such document graphs are discussed in detail in (Mani and Bloedorn 1997a), (Mani and Bloedorn 1997b), so we will merely list them here. They include MITRE's Alembic part-of-speech tagger (Aberdeen et al. 1995) (used to extract phrases by means of finite-state patterns over part-of-speech tags), proper name extraction using SRA's NameTag (Krupka 1995), and extraction of noun synonyms and hypernyms via WordNet (Miller 1995).

Computing Salience based on Cohesion

Given this graph, a variety of different methods can be used to compute salience. The first method is a baseline method, which ignores most of the information in the graph, using repetition alone.

Method 1: tf.idf In Method 1, we weight words based on repetition alone, using the tf.idf weighing metric. The weight dw_{ik} of term k in document i is given by:

$$dw_{ik} = tf_{ik} * (ln(n) - ln(df_k) + 1) \quad (1)$$

where tf_{ik} = frequency of term k in document i, df_k = number of documents in a reference corpus (derived from the TREC collection (Harman 1994)) in which term k occurs, n = total number of documents in the reference corpus.

In this method, the weight of a clause is simply the sum of the tf.idf weights of the words in the clause.

Method 2: Spreading In Method 2, the cohesion links in the graph are used to compute salience, using a spreading activation search algorithm. The spreading activation search activates words related to initial words, or topics. Given a topic that captures a user interest, the refinement phase of processing begins by computing a salience function for text items based on their relation to the topic. The algorithm (derived from (Chen, Basu, and Ng 1994)) is used to find nodes in graph related to topic nodes. (If there is no topic, nodes in initial positions are chosen as topic nodes based on a tf.idf threshold.)

The method, which corresponds to a strict best-first search of the graph, begins by pushing the nodes matching the given query terms onto an input stack, which is kept sorted by decreasing weight. (The matching uses stemming based on (Porter 1980).) The method then iterates until a terminating condition is reached, popping the current node off the stack and placing it on the output, and then finding successor nodes linked to the current node in the graph and placing them on the stack. The weight of a successor node is a function of the source node weight and the link type weight. Each different link type has a dampening effect on the source node weight. Since the graph may have cycles, the weight of a successor node is determined by the strongest path to the node. Thus, the successor node weight is the maximum of its new weight and its old weight. The algorithm terminates when the slope change of the total activation weight of the graph drops below a fixed delta. The termination condition examines a window of the last 40 nodes and stops if the standard deviation of slopes in that window is less than 0.1.

The spreading activation is constrained so that the activation decays by link type and text distance. We use the following ordering of different link types, with earlier links in the ordering being heavier (and thus

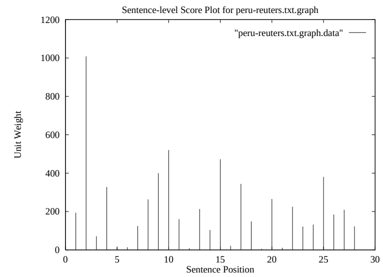

Figure 2: Sentence-level Activation Weights from Raw Graph (Reuters news)

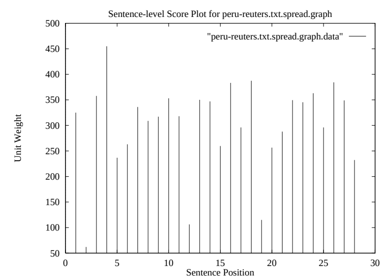

Figure 3: Sentence-level Activation Weights from Spread Graph (Reuters news; topic: *Tupac Amaru*)

having less dampening effect) than later links:

SAME COREFERENCE NAME PHRASE ALPHA ADJ

For ADJ links, successor node weight is an exponentially decaying function of current node weight and the distance between nodes. Here distances are scaled so that traveling across sentence boundaries is more expensive than traveling within a sentence, but less than traveling across paragraph boundaries. For the other link types, the successor weight is the product of link weight and current node weight.

As an example, we show a sentence-level plot of the activation weights for a Reuters article, where the weight at a given sentence position is calculated as the average its constituent word weights in Figure 2. The results after spreading given the topic *Tupac Amaru*, are shown in Figure 3. The spreading has changed the activation weight surface, so that some new related peaks (i.e., local maxima) have emerged (e.g., sentence 4), and old peaks have been reduced (e.g., sentence 2, which had a high tf.idf score, but was not related to *Tupac Amaru*). The exponential decay function is also evident in the neighborhoods of the peaks.

This spreading activation technique has been demonstrated in a number of intrinsic and extrinsic summarization evaluation experiments to be effective in summarization, in particular helping to establish what is salient in relation to a topic (Mani and Bloedorn 1997a), (Mani and Bloedorn 1997b). Since this method was originally intended for topic-sensitive summaries (in particular for generating multi-document topic-sensitive summaries), it requires a specific topic (or default) to help determine the starting nodes.

Using this technique, the weight of a clause is determined by the spreading activation given the initial tf.idf weights and the topic terms.

Method 3: Local Weighting In Method 3, the cohesion relations in the graph are again used to compute salience, but unlike Method 2, the computed salience weights are not spread through the graph. Instead, the more strongly connected the node is, the higher the weight. The weight of a word (instance) node is simply the sum of the link weights to that node. The weight of a clause, in this scheme, is thus the sum of all the weights of words in that clause. Again, nodes are initially weighted with tf.idf weights. Note that this method does not require any initial topic to start with.

Coherence

Cohesion is often contrasted with coherence, which, as mentioned earlier, has to do with macro-level, deliberative structuring of multi-sentence text in terms of relations between sentences and clauses. A variety of different theories from a wide variety of intellectual disciplines have been proposed to provide an analysis of argumentation structure in multi-sentence text, including Rhetorical Structure Theory (RST) (Mann and Thompson 1988), Discourse Grammar (Longacre 1979), Macrostructures (Van Dijk 1988), Coherence Relations (Hobbs 1985), Toulmin Structure (Toulmin 1958) (the latter is focused on argumentation, not text), etc. The diversity of approaches and the lack of formal specification of the semantics of the coherence relations may seem somewhat discouraging, but a note of encouragement can be found in the observation by (Hovy 1990) that it is possible to group the more than 350 different relations proposed in the literature in terms of a hierarchy based on 16 core relations. Another encouraging trend is found in the development by (Knott and Dale 1996) of an empirical approach for determining coherence relations based on a taxonomy of cue phrases. Finally, Daniel Marcu (Marcu 1996), (Marcu 1997b) has developed techniques for automatic parsing of RST-structure, based on an analysis of cue phrases in a large corpus. This method disambiguates between sentential and discourse uses of cue phrases, reporting precision of 80.8% and recall of 89.5% in identifying discourse uses. (Of course, further studies need to be undertaken to assess what sorts of errors occur with this method, as well as to how well it scales up to other genres).

In this paper, we will use a corpus of 5 texts already marked up by hand with RST-structure, supplied to us by Marcu along with ground-truth judgments of salience (to be discussed below), a sixth text (Stargazer, used by Marti Hearst in her work on segmenting text with cohesion relations using the Text Tiling algorithm (Hearst 1994)) marked up with RST-structure by Marcu's parser. All six of these texts are expository texts; to include another genre, we also include a seventh text which is a news story (Peru), where we marked up the RST-structure ourselves based on the news schema of (Van Dijk 1988). Note that human determination of salience was available only for texts 1-5.

Computing Salience based on Coherence

To maximize points of comparison, we use the same salience metric for coherence described by Marcu (Marcu 1997b), (Marcu 1997a). The terminals in a RST tree are individual clauses, whereas all the other nodes span more than one clause. The RST-trees in his scheme are strict binary trees. The children of a parent are either both nucleus nodes (in which case the node represents a paratactic relation relating the two children), or one is a nucleus and the other is a satellite (the node represents a hypotactic relation). In determining salience, the crucial point is that a nucleus is considered more salient than a satellite. The coherence salience function is then simply one of tree-depth in the RST tree. Each parent node identifies its nuclear children as salient, and this identification is continued recursively down the tree. A partial ordering is defined on these salient clauses, based on depth from the root: the salient clauses identified with a parent are higher in the ordering than salient clauses identified with its children.

Combining Cohesion and Coherence

Comparing Salience

We now compare salience scores produced by the cohesion and coherence methods and compare them with ground truth. The ground-truth judgments for clause salience in texts 1-5 were carried out in research by (Marcu 1997a). In Marcu's work, 13 judges each assigned every clause in the 5 texts a score of 2, 1, or 0, where 2 means the clause should appear in a concise summary, 1 means that it should appear in a long summary, and 0 means that it is unimportant. Now (the procedure departs somewhat from Marcu's) we assign each clause a salience score which is the sum of its scores across all judges. For each text t we define a cutoff n_t as 20% (for comparison, we also use 10%) of the total number of clauses in the text. For each text t, we determine the top n_t salient clauses ranked by the human, the top n_t computed by coherence, and the top n_t computed by each of the cohesion methods. The top clauses retrieved in these cases are shown in Table 1. As can be seen, on the mars text, the spreading picks up two of the four 20% salient clauses, predicting correctly their relative ordering, as does the coherence method. On the cars text, spreading picks out four of the nine salient clauses, predicting correctly their relative ordering. The coherence picks out the same number, but does not predict correctly the ordering. This suggests the cohesion methods in certain

Human	tfidf	spread	local	coherence
(1) 4.12	13.18	4.3	5.9	4.[10.11.12]
4.12.5.18	13.18.6.12	4.3.9.12	5.9.8.7	4.[10.11.12]
(2) 4.1.15	9.25.14	15.8.9	11.10.6	6.[7.9.10.11.12]
4.1.15.8.12.[5.9.10.11]	9.25.14.1.10.3	15.8.9.16.10.3	11.10.6.7.24.15	6.[7.9.10.11.12]
(3) 11.13	10.11	12.11	4.7	2.13
11.13.4.[2.6.17]	10.11.15.13	12.11.8.9	4.7.8.12	2.13.[6.17.12.20]
(4) 24.4.6	23.1.16	19.23.1	21.20.24	1.[16.17.18.21.22.23.24]
24.4.6.3.17.[5.16]	23.1.16.19.24.3	19.23.1.16.17.24	21.20.24.23.22.19	1.[16.17.18.21.22.23.24]
(5a) 2.19.10.61.5	18.19.63.24.37	18.59.56.37.48	59.31.22.66.65	64.13.16.2.61
.13.[37.63]	.44.17	.42.63	.43.57	.[1.19.37.48.49.51.62.63.64]
(5b) 2.19.10.61.5	18.19.63.24.37	18.59.56.37.48	59.31.22.66.65	64.13.16.2.61
.13.[37.63].21.62	.44.17.41.12.55	.42.63.39.68.7	.43.57.51.28.35	.[1.19.37.48.49
.25.49.7.[35.66]	.51.42.56.48	.19.55.14.13	.60.49.5.53	.51.62.63.64.66]
(6a)	5.12.41.40.25	25.24.35.34.61	32.[59.53].105	[0.2.4.12.16.21.25.27.28.29.30.47
	.35.[24.18].61.34	.18.20.6.21.8	.[7.73].36.6.84.44	53.55.57.61.62.64.65.73.79.81.89.90]
(6b)	5.12.41.40.25	25.24.35.34.61	32.[59.53].105.[7	[0.2.4.12.16.21
	.35.[24.18].61.34	.18.20.6.21.8	.73].36.6.84.44	25.27.28.29.30.47
	.6.20.9.71.10	.62.9.10.38.11	.40.[87.41].35.[56	53.55.57.61.62.64
	.31.57.81.60.8	.57.31.58.30.19	.38.22.17].24.12	65.73.79.81.89.90]
(7a)	33.16.12.11	4.34.3.33	[18.16].20.24	1-4
(7b)	33.16.12.11.18.23.10.20	4.34.3.33.15.9.16.12	[18.16].20.24.10.26.28.11	1-4.5-8

Table 1: Ordering of top clauses for different texts computed by coherence and cohesion methods and human judgments. Text names/topics (the latter created by hand for spreading) are: (1) Mars/mars, (2) cars/options, (3) pop-top/rivet, (4) smart-cards/card, (5) aerobics/metabolic (6) stargazer/telescope, moon, earth, (7) peru/Tupac Amaru. The first row for each text 1-4 shows 10% cutoff, the second row 20% cutoff. In texts 5-6, the scores for each text are split across multiple rows for reasons of space. For (5), the first two rows (marked (5a)) correspond to the 10% cutoff, the next three rows, marked (5b), correspond to the 20% cutoff. Clauses in [ ] are equivalent as determined by the total score given by the 13 human judges, or by level in the RST tree.

cases are quite helpful in picking out salient clauses. To examine it more precisely, the precision and recall (because we are cutting off at a fixed percentage, precision = recall) of these different methods against the human judgments of salience, where available, is shown in Table 2.

As can be seen from Tables 1 and 2, the spreading method outperforms the other cohesion methods in accuracy at both 10% and 20%. Further, the local weighting cohesion method is actually worse than the baseline tf.idf method. The coherence beats the spreading on both cutoffs, but at 10%, both perform poorly, although the difference between them is somewhat less. These results show that although the coherence method beats the cohesion methods in accuracy of determining clause salience, the best cohesion method can reach 76% of the accuracy levels of the coherence method in determining salience. While the absolute accuracy of the spreading method is still quite low (45.4), it is worth pointing out that that the cohesion methods fare best on texts where the different cohesion link types can be exploited. The paucity of proper name anaphora and even synonym/hypernym links in these texts is responsible for much of the low performance, which is based on adjacency and repetition links alone.

We also computed the Spearman rank-order correlation coefficient between cohesion methods for salience and human judgments of salience. Although we obtained a significant positive correlation ($p < 0.05$) between the tfidf-based (i.e., baseline) salience ordering and the human judges' ordering for one text (smart-cards), the other two cohesion methods each yielded significant positive correlations ($p < 0.05$) with the human salience ordering for the cars, mars, and pop-top texts. Since the cohesion methods are highly robust and knowledge poor, the results are encouraging.

Comparing Structure

We have seen so far that the cohesion methods for computing clause salience, while they are somewhat less accurate than coherence methods, are significantly positively correlated with human judgments of salience. This suggests that both cohesion and coherence are useful, in the analysis and refinement phases of summarization, in determining what is salient. Now, while salience can be used to extract material, it is useful also to represent any implicit discourse-related structure discovered in the text, which can be used in particular in the synthesis phase of summarization to generate abstracts. We now compare the types of structure that coherence and cohesion can discover.

In general, one suspects that cohesion is useful for coarse-grained (e.g., paragraph-sized) segmentation of topic structure, whereas coherence is more oriented towards fine-grained (clause-sized) segmentation of topic structure. An inspection of the performance of (our implementation (Mani et al. 1997) of) the Text Tiling algorithm of (Hearst 1994) against texts 1-5 confirms this. The Tiling algorithm created segments (tiles) very close to paragraph boundaries in each of those texts. One would expect the paragraph-breaks to be higher in the RST-tree than non-paragraph breaks, and this is borne out by examination of texts 1-5 and their RST-trees.

Given this, one might expect that cohesion can help with attachment of RST structure, where the RST structure spans or needs to span a major topic boundary. We therefore intend to examine cases where the

Text	tfidf	spread	local	coherence
Mars	0	50	0	100
	50	50	25	50
Cars	33	33	0	0
	33	66	50	33
Pop-top	50	50	50	50
	50	25	25	100
Smart-cards	0	0	33	33
	50	50	17	50
Aerobic	43	29	0	60
	21	36	29	64
Average	25.2	32.4	16.6	48.6
	35.8	45.4	29.2	59.4

Table 2: Precision and Recall of Salient Clauses by coherence and cohesion methods. The first row for each text 1-4 shows 10% cutoff, the second row 20% cutoff.

Text	Size	Cohesion		Coherence	
		Nodes	Leaf Size	Nodes	Leaf Size
mars	161	52	2.7	32	8.9
cars	257	43	3	46	10.3
pop-top	225	25	10.8	37	11.4
smart-cards	295	11	23	50	10.9
aerobic	726	155	3.2	136	10.5
stargazer	1924	519	4.1	225	17.0
peru	632	33	13.7	58	18.0
Average	602.8	128.3	8.6	83.4	12.4

Table 3: Size differences among trees generated by Cohesion and Coherence methods. **Size** indicates the length of the text in words, **Nodes** indicates the number of nodes in the tree, and **Leaf Size** indicates the average number of words at a leaf.

RST-structure might be hard to extract automatically (for example, rhetorical relations not explicitly marked in the text, or those marked with highly ambiguous markers like "and"), and to investigate whether cohesion-based similarity can be used to make attachment decisions. Of course, such attachment may not be able to provide labels for the cohesion relations.

Now, it is of course possible to use cohesion alone to extract structure from the text. One expects this type of structure (which is not based on analysis of cue phrases) to be quite different from RST structure, as it is oriented towards discovering structural boundaries based on connectedness. Nevertheless, it is worthwhile comparing the two structures to examine whether there are any points of commonality. To this end, we have investigated the use of agglomerative clustering (Yaari 1997) with (local weighting) cohesion to induce hierarchical tree-structure from text. This algorithm, given a set of segments which are elementary text units, combines together the most similar pair of adjacent segments into a single segment, and recurses until it has merged all the segments. Here, we apply this algorithm by beginning with word instances in the graph, and merging them together based on how strongly connected the words are, via local cohesion weighting. The (binary) trees induced by this cohesion method for the 7 texts can then be compared with the RST trees.

In Table 3, we compare the cohesion and RST trees in terms of size. As can be seen, coherence has much less variability in leaf size, since the minimal units are usually clauses. It is possible to start with larger minimal segments in the cohesion method, such as clauses or paragraphs, but the goal here was to let the data drive the minimal segment size, rather than fixing the unit. Also, the cohesion-based stargazer tree is relatively much larger than the coherence one. In the future, we may attempt to constrain the number of nodes in the cohesion trees to make them smaller or comparable in size to the RST trees.

If we do a more detailed comparison of the trees, we find many differences, but also, occasionally, some points of similarity. For example, in the pop-top text, although the two structures cross branches in numerous places, ten pairs of nodes, five of which are nuclei, have similar spans across the two trees. Figure 4 contains pop-top subtrees in which three pairs of internal nodes (1-3, 14-16, and 15-16) have similar spans. Both methods identify the introduction (clauses 1-3) as a distinct node, although cohesion gives its subtree a left-branching structure, while coherence gives it a right-branching structure; and both methods identify the join (clauses 14-16), with the same internal structure. In other texts, however, the similarities are much fewer.

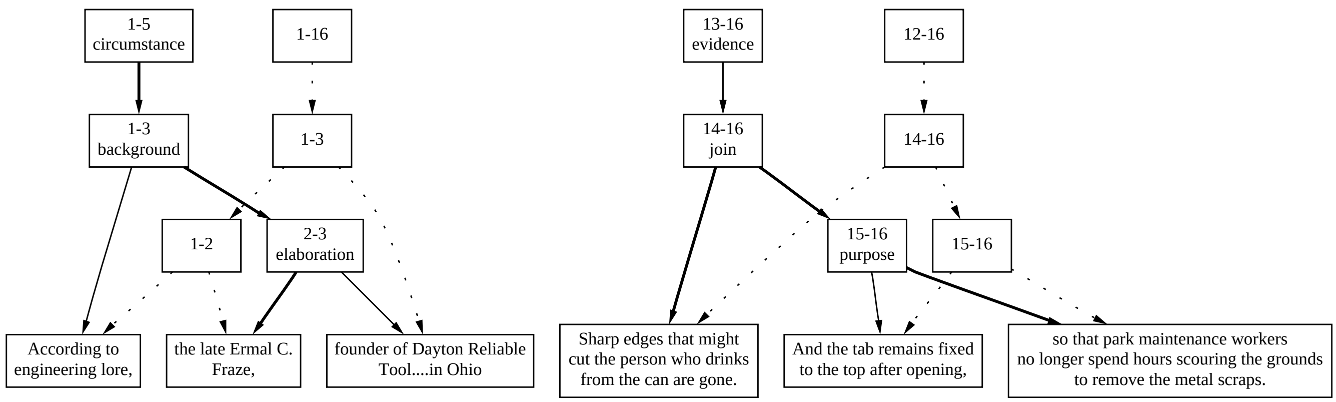

Figure 4: Some RST (plain edges) and Cohesion (dotted edges) subtrees for pop-top text. Bold edges point to nuclei. Ranges show clause spans.

Conclusion

In any summarization task, a key function is to determine a salience ordering for text units. We have explored three different cohesion methods for computing salience, and compared these methods against a coherence-based method in terms of accuracy with respect to human judgements of salience. Despite the lack of links based on proper name anaphora and synonymy in these texts, the results of the spreading-based cohesion method, while less accurate overall than the coherence method, are encouraging. Further, a method for discovering discourse-related text structure based on local cohesion was introduced. We also compare such trees with coherence-based trees. In the future, we expect to directly explore the *combination* of coherence and cohesion-based methods to compute salience and to discover important discourse-related structural units in text.

Acknowledgments

We are deeply indebted to Daniel Marcu, for providing both a corpus of RST-annotated texts, as well as human judgments of these. His innovative approach to the exploitation of RST has inspired much of this research.

References

Aberdeen, J., Burger, J., Day, D., Hirschman, L., Robinson, P., and Vilain, M. 1995. MITRE: Description of the Alembic System Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, Maryland, November.

Aone, C., Okurowski, M. E., Gorlinsky, J., and Larsen, B. 1997. "A Scalable Summarization System using Robust NLP", in Mani, I., and Maybury, M., eds., Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, pp. 66-73.

Barzilay, R., and Elhadad, M. 1997. "Using Lexical Chains for Text Summarization", in Mani, I., and Maybury, M., eds., Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, pp. 10-17.

Boguraev, B., and Kennedy, C. 1997. "Salience-based Content Characterization of Text Documents", in Mani, I., and Maybury, M., eds., Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, pp. 2-9.

C.H. Chen, K. Basu, and T. Ng. 1994. An Algorithmic Approach to Concept Exploration in a Large Knowledge Network. Technical Report, MIS Department, University of Arizona, Tucson, AZ.

Edmundson, H.P. 1969. "New methods in automatic abstracting", Journal of the Association for Computing Machinery, 16, 2, pp. 264-285.

Green, S.J. 1997. "Building Hypertext Links in newspaper articles using Semantic Similarity", Technical Report, Department of Computer Science, University of Toronto.

Halliday, M. and Hasan, R. 1996. "Cohesion in Text", London, Longmans.

Donna Harman (editor). 1994. *An Overview of the Third Text Retrieval Conference*, National Institute of Standards and Technology, NIST Special Publication 500-225, Gaithersburg, MD.

Hearst M. 1994. "Multi-Paragraph Segmentation of Expository Text", Proceedings of ACL-94, Las Cruces, New Mexico.

Hobbs, J.R. 1985. "On the Coherence and Structure of Discourse", Technical Report No. CSLI-85-37, Center for the Study of Language and Information, Stanford University.

Hovy, E. 1990. "Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations", in Proceedings of the Fifth International Workshop on Natural Language Generation, Dawson, Pa.

Knott, A. and Dale, R. 1996. "Choosing a Set of Coherence Relations for Text Generation: a Data-

Driven Approach", in M. Zock, ed., Trends in Natural Language Generation, An Artificial Intelligence Perspective, Springer-Verlag.

George Krupka. 1995. SRA: Description of the SRA System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, Maryland, November.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A Trainable Document Summarizer. In *Proceedings of ACM-SIGIR'95*, Seattle, WA.

Liddy, E.R. 1991. "The discourse-level Structure of Empirical Abstracts: An Exploratory Study", Information Processing and Management, 27, 1, 55-81.

Longacre, R. 1979. "The paragraph as a Grammatical Unit', in T. Givon, ed., Syntax and Semantics, 12, Academic Press.

Mani, I., and Bloedorn,. E. 1997. "Summarizing Similarities and Differences Among Related Documents", Proceedings of RIAO-97, Montreal, Canada, pp. 373-387.

Mani, I., and Bloedorn, E. 1997. "Multi-document Summarization by Graph Search and Merging", Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97), Providence, RI, pp. 622-628.

Mani, I., House, D., Maybury, M., and Green, M. 1997. "Towards Content-Based Browsing of Broadcast News Video", in Maybury, M., ed., Multimedia Information Retrieval, AAAI/MIT Press.

Mann, W. C. and Thompson, S, A. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8, 3, pp. 243-281.

Marcu, D. 1996. "Building Up Rhetorical Structure Trees", Proceedings of AAAI-96, Portland, Oregon, pp. 1069-1074.

Marcu, D. 1997. "From discourse structures to text summaries", in Mani, I., and Maybury, M., eds., Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, pp. 82-88.

Marcu, D. 1997. "The Rhetorical Parsing of Natural Language Texts", Proceedings of ACL/EACL'97, Madrid, Spain.

Kathy McKeown and Dragomir Radev. 1995. Generating Summaries of Multiple News Articles. In *Proceedings of ACM-SIGIR'95*, Seattle, WA.

Miike, S., Itoh, E., Ono, K., and Sumita, K. 1994. "A Full-Text Retrieval System with a Dynamic Abstract Generation Function", Proceedings of ACM-SIGIR'94, Dublin, Ireland.

George Miller. 1995. WordNet: A Lexical Database for English. In *Communications of the ACM*, 38, 11, 39-41.

Morris, J., and Hirst, G. 1991. "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", Computational Linguistics, 17, 1, pp. 21-43.

Paice, C. and Jones, P. 1993. "The Identification of Important Concepts in Highly Structured Technical Papers", Proceedings of ACM-SIGIR'93, Pittsburgh, PA.

Porter, M.F. 1980. "An Algorithm For Suffix Stripping", Program, 14, 3, July 1980, pp. 130-137.

Toulmin, S. 1958. The Uses of Argument, Cambridge University Press.

Van Dijk, T. A. 1988. "News as Discourse", Lawrence Erlbaum, Hillsdale, NJ.

Yaari, Y. 1997. "Segmentation of Expository Texts by Hierarchical Agglomerative Clustering", Technical Report, Bar-Ilan University, Israel. Also available as cmp-lg/9709015.