CST Bank: A Corpus for the Study of Cross-document Structural Relationships

Dragomir Radev*†, Jahna Otterbacher*, Zhu Zhang*†

*School of Information
†Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI USA
{jahna,radev,zhuzhang}@umich.edu

Abstract

Clusters of multiple news stories related to the same topic exhibit a number of interesting properties. For example, when documents have been published at various points in time or by different authors or news agencies, one finds many instances of paraphrasing, information overlap and even contradiction. The current paper presents the Cross-document Structure Theory (CST) Bank, a collection of multi-document clusters in which pairs of sentences from different documents have been annotated for cross-document structure theory relationships. We will describe how we built the corpus, including our method for reducing the number of sentence pairs to be annotated by our hired judges, using lexical similarity measures. Finally, we will describe how CST and the CST Bank can be applied to different research areas such as multi-document summarization.

1. Introduction

Multiple news stories on the same event present some challenges for natural language processing. They contain similar information and yet they also exhibit a number of interesting properties: paraphrases, partial agreement, differences in judgment and emphasis, and even contradiction. When news sources are tracked over time, more phenomena can be observed: updates, corrections, etc. For example, Figures 1 and 2 show two articles from different news agencies about the same story - the crash of a small plane into the tallest skyscraper in Milan. Some observations that can be made include:

- Sentences 1:9 and 2:2 both discuss the casualties in the incident. However, sentence 2:2 elaborates on 1:9.
- Sentences 1:2 and 2:1 contradict each other (25th floor vs. 26th floor).
- Sentences 1:1 and 2:1 contain the same information but sentence 2:1 attributes this information to a source while 1:1 presents it as a fact.
- Sentences 2:5 and 2:6 present some historical background about the event.

In (Radev, 2000), we proposed Cross-document Structure Theory (CST), a functional theory for multi-document discourse structure. CST is used to describe semantic connections among units of related documents such as "elaboration", "contradiction", "attribution", and "historical background" as illustrated above. CST is related to RST (Mann and Thompson, 1988) but assumes no deliberateness of writing and no underlying tree representation. While the graph-like representation looks like "semantic hyperlinks", the relationships are all linguistically motivated.

The proposed taxonomy for CST relationships was first described in (Radev, 2000).[1] It should be noted that some CST relationships, such as *identity*, are symmetric (*binuclear* in RST terms), while some other ones, such as *subsumption*, do have directionality, i.e., they have *nucleus* and *satellite*. Some of the relationships are direct descendents of those used in SUMMONS (Radev and McKeown, 1998), however, in CST, all relationships are domain-independent.

In this paper we will discuss the creation of the first CST Bank, a corpus of document clusters manually annotated for CST relationships. We will also present how lexical similarity between sentences was used as a way to narrow down the list of possible candidate sentence pairs.

2. CST Bank Composition

The first phase of CST Bank includes six clusters of related news articles from various sources. The clusters were chosen to be diverse with respect to their topics, the time span across the documents, the cluster size, and the news agencies from which the articles were collected. Figure 3 shows the characteristics of the clusters. The cluster names reflects the source from which the cluster of documents was obtained.

Three of the clusters were collected from secondary sources while three were collected by the authors. The DUC cluster was obtained from the 2001 Document Understanding Conference (DUC) training data, the HKNews cluster was taken from the Hong Kong Corpus, and the Novelty cluster was a cluster from the 2002 TREC Novelty track test data. The Milan 9 and Gulfair 11 clusters were collected by the authors live from the Web from several news sites: USA Today, MSNBC, CNN, FOX News, the BBC, the Washington Post and ABC News. Finally, the NIE cluster was collected automatically using the NewsTroll agent of NewsInEssence (http://www.newsinessence.com).

The first CST Bank cluster, Milan 9, was used strictly for training and corpus development purposes. It was annotated for CST relationships by two of the authors in developing the markup scheme and the guidelines to be used by the independent judges to be hired. In phase one, the five clusters that were annotated by the judges were DUC, Gulfair 11, HKNews, NIE, and Novelty.

[1] It can also be found at $http://tangra.si.umich.edu/clair/CSTBank/taxonomy.pdf$ in its most recent form.

Plane hits skyscraper in Milan

(1) A small plane has hit a skyscraper in central Milan, setting the top floors of the 30-story building on fire, an Italian journalist told CNN. (2) The crash by the Piper tourist plane into the 26th floor occurred at 5:50 p.m. (1450 GMT) on Thursday, said journalist Desideria Cavina. (3) The building houses government offices and is next to the city's central train station. (4) Several storeys of the building were engulfed in fire, she said. (5) Italian TV says the crash put a hole in the 25th floor of the Pirelli building, and that smoke is pouring from the opening. (6) Police and ambulances are at the scene. (7) Many people were on the streets as they left work for the evening at the time of the crash. (8) Police were trying to keep people away, and many ambulances were on the scene. (9) There is no word yet on casualties.

Figure 1: Milan article 1 (CNN).

Plane Slams Into Milan Skyscraper

(1) A small plane crashed into the 25th floor of a skyscraper in downtown Milan today. (2) At least three people, including the pilot, were dead, Italy's ANSA wire service said. (3) Dozens of people in the Pirelli building were injured after several floors of the 32-story building caught fire, local reports said. (4) Only the pilot was on board the plane, reported The Associated Press. (5) It was the second time since the Sept. 11 terror attacks that a plane has struck a high-rise building, and the crash raised fears of another attack. (6) On Jan. 5, a 15-year-old boy crashed a stolen plane into a building in Tampa, Fla.

Figure 2: Excerpts from Milan article 2 (ABCNews).

3. Similarity Metrics

Building CST Bank involves asking human annotators to mark CST relationships in the document clusters. Human annotation is not only expensive, but also hard to achieve agreement on, due to the inherently ambiguous nature of natural language. The large search space makes the situation even worse. In a ten-document cluster with 20 sentences on average in each document, for example, a human judge will have to examine roughly 20,000 sentence pairs if he or she wants to exhaust all possibilities. This is an incredibly tedious job in any sense, and because of that, it is very difficult for multiple judges to reach reasonable agreement on the annotation.

One possible way to alleviate the problem is to exploit the observation that CST relationships are unlikely to exist between sentences that are *lexically* very dissimilar to each other. In other words, certain similarity measures might behave as a useful proxy for finding CST-related sentence pairs.

To test the hypothesis, we experimented with the following similarity metrics and measured their correlation with CST-relatedness:

- Word-based cosine similarity

$$cos(S_1, S_2) = \frac{\sum s_{1,i} * s_{2,i}}{\sqrt{\sum (s_{1,i})^2} * \sqrt{\sum (s_{2,i})^2}}$$

- Word overlap

$$wol(S_1, S_2) = \frac{\#CommonWords(S_1, S_2)}{\#Words(S_1) + Words(S_2)}$$

- Longest common subsequence

$$lcs(S_1, S_2) = \frac{\#Words(LCS(S_1, S_2))}{\#Words(S_1) + \#Words(S_2)}$$

- The Bleu metric (Papineni et al., 2002) which is a linear combination of n-grams with a penalty for length mismatch.

4. Similarity Approximation

4.1. Step 1: Bootstrapping from a Very Small Corpus

We first started with a very small corpus (MI3), which contains 3 articles and 45 sentences in total. Two coauthors went through the corpus very carefully and identified 16 CST pairs. After one more pass of negotiation and discussion, the two judges reached agreement on all pairs except 2, which they labeled as different CST types.

With the annotated CST relationships in MI3, we can now measure the performance of various similarity metrics. Figure 4 summarizes the results. Recall and precision are used to measure agreement between the automatic similarity identification and the manual golden standard. We also present the number of candidate pairs suggested by each metric.

We observe the following patterns from the data:

- Cosine similarity is overall the best of the four metrics in terms of capturing CST-relatedness between sentence pairs, but is not necessarily best for our purpose because it tends to report too many potentially "interesting" pairs.
- Bleu is the tightest similarity measure but also the least robust one, because it measures overlap for up to 4-grams between sentence pairs. CST-related sentences typically don't have many 4-grams in common.
- "Word overlap" is the most appropriate similarity measure for our purpose in terms of providing both reasonable recall and not very many sentence pairs. We ended up choosing a word overlap threshold of 0.12 (abbreviated as WO-0.12) for later experiments, which gives us 87.5% recall on MI3 while not presenting too many candidate sentence pairs.
- Excluding stop words doesn't help in general.

Cluster	Topic	Articles	Time span	Ave. length (sent.)	No. sources	Clustering method
Milan 9	Milan plane crash	9	2 days	30	5	manual
DUC	John Lennon biography	4	4 years	46	4	manual
Gulfair 11	Bahrain plane crash	11	4 days	27	6	manual
HKNews	Air and water quality	8	2.5 years	32	1	manual
NIE	N. Korea nuclear weapons	5	18 days	14	3	automatic
Novelty	Cancer and power lines	4	4 years	21	2	manual

Figure 3: Characteristics of CST Bank Phase I Clusters

	Threshold	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
Cosine	Recall	68.75%	56.25%	43.75%	43.75%	25.00%	18.75%	0.00%	0.00%	0.00%	0.00%
	Precision	8.09%	9.78%	12.50%	21.88%	22.22%	37.50%	0.00%	0.00%	0.00%	0.00%
	#Pairs	136	92	56	32	18	8	2	0	0	0
Word overlap	Recall	68.75%	37.50%	18.75%	6.25%	6.25%	0.00%	0.00%	0.00%	0.00%	0.00%
	Precision	14.47%	30.00%	37.50%	25.00%	50.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	#Pairs	76	20	8	4	2	0	0	0	0	0
LCS	Recall	62.50%	31.25%	12.50%	6.25%	6.25%	0.00%	0.00%	0.00%	0.00%	0.00%
	Precision	14.29%	27.78%	33.33%	25.00%	50.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	#Pairs	70	18	6	4	2	0	0	0	0	0
BLEU	Recall	31.25%	31.25%	18.75%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	Precision	33.33%	35.71%	60.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	#Pairs	15	14	5	0	0	0	0	0	0	0

Figure 4: Performance of various similarity measures on MI3 (without stop word removal)

4.2. Step 2: Experimenting with the Larger Corpus

In step 2, we ran WO-0.12 on a bigger corpus, MI9, which contains 9 articles, 269 sentences in total. It produced 1815 candidate sentence pairs (out of 18,023 possible pairs). The two coauthors again went through them and identified 1,145 CST-related pairs.

With the annotated data on MI9, we reran the similarity experiments on MI9. The results are summarized in figure 5.

Cluster	Judges	No. sentence pairs
DUC	C,D,E	475
Gulfair 11	B,F	2,242
HKNews	G,H	1,729
NIE	A,D,E	421
Novelty	A,B,F	64

Figure 6: Judges annotating CST Bank clusters

5. CST Bank Data Collection

Eight judges were hired for the annotation of the first five clusters of CST Bank. Two one-hour training sessions were held and each judge attended one of the sessions. During the training, the authors explained the motivation of CST relationships and their definitions, discussed how the various CST relationships differ from one another, and presented some examples of annotated sentence pairs. Most importantly, the judges were given the CST Bank Annotation Guidelines, which will be discussed in more detail in the next section. Each section of the guidelines includes 15 practice sentence pairs. The judges were asked to complete these examples in order to ensure that they had gained sufficient understanding of the annotation process and the definitions of the relationships.

Once the judges had completed the training session, they were assigned an annotation packet, which contained hard copies of the source articles and the similar sentences pairs to be annotated for a particular cluster. Clusters were assigned one at a time in order to ensure the quality of the annotations. Five of the eight judges requested additional packets to annotate. Figure 6 shows each cluster, the judges who annotated it, and the number of similar sentence pairs that were presented to the judges.

Since CST relationships are not mutually exclusive, and judges often assign more than one CST relationship to a given sentence pair, it should be noted that the judges do not always agree. In addition, this makes it difficult to quantify the level of interjudge agreement between them. Therefore, we have found the level of agreement with respect to the existence of CST relationships (regardless of type). Kappa (Carletta, 1996) for the corpus is 0.53.

6. Annotation Guidelines

As mentioned previously, two of the authors independently annotated the development cluster, MI9, in order to understand the difficulties that might come up during the annotation process. In addition, we were able to identify many examples of tricky sentence pairs for inclusion in the CST Bank Annotation Guidelines. The guidelines are meant to serve as instructions for the judges as well as for researchers who wish to use CST Bank data in their work.[2]

The guidelines include a section on the motivation for CST relationships, as well as their definitions and examples of sentence pairs for each of the eighteen relationships that have been identified thus far. We have alluded to the fact that CST relationships might be used in multi-document summarization, in order to improve the quality of summaries. However, we have not provided specific information about this, as we imagine that CST relationships might be useful for a number of different problems and tasks, and in addition, they can highlight interesting properties of evolving news stories.

The largest sections of the guidelines are devoted to providing explicit examples and discussion of sentence

[2]The annotation guidelines are available at $http://tangra.si.umich.edu/clair/CSTBank/annotation\_guide.pdf$.

	Threshold	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
Cosine	Recall	72.58%	59.52%	45.46%	35.29%	23.93%	20.04%	15.85%	12.86%	11.37%	8.87%
	Precision	29.70%	41.12%	60.32%	78.49%	86.02%	90.13%	91.38%	96.99%	99.13%	100.00%
	#Pairs	2451	1452	756	451	279	223	174	133	115	89
Word overlap	Recall	63.61%	29.31%	20.54%	17.55%	14.96%	12.36%	11.37%	10.47%	9.27%	8.47%
	Precision	68.45%	93.63%	97.17%	99.44%	99.34%	99.20%	99.13%	99.06%	100.00%	100.00%
	#Pairs	932	314	212	177	151	125	115	106	93	85
LCS	Recall	56.93%	27.52%	18.84%	17.05%	14.86%	12.26%	11.27%	10.17%	9.07%	8.47%
	Precision	67.26%	94.52%	96.92%	99.42%	99.33%	99.19%	99.12%	99.03%	100.00%	100.00%
	#Pairs	849	292	195	172	150	124	114	103	91	85
BLEU	Recall	23.13%	19.74%	16.10%	14.36%	12.96%	11.17%	9.47%	8.87%	7.88%	7.68%
	Precision	79.86%	87.80%	90.48%	99.31%	99.24%	99.56%	100.00%	100.00%	100.00%	100.00%
	#Pairs	291	226	179	145	131	113	95	89	79	77

Figure 5: Performance of various similarity measures on MI9 (without stop word removal)

pairs taken from the MI9 training cluster. Another section focuses on explaining some of the subtle differences between relationships such as follow up versus elaboration/refinement and description versus historical background, which can often be confused when first encountering them. Finally, another important feature of the annotation guidelines is the inclusion of fifteen pairs of practice sentences. As mentioned in the last section, we used these practice sentences in our training sessions with the eight judges.

7. Related Work

This paper builds on earlier work on rhetorical structure analysis and sentence similarity identification. (Hatzivassiloglou et al., 2001) present a statistical similarity measuring and clustering tool, SIMFINDER, that organizes small pieces of text from one or multiple documents into tight clusters. More specifically, they use logistic regression models for classifying similar/dissimilar paragraphs. Both primitive and composite features are used in their model. The problem is in some sense simpler than ours, because we are identifying much more fine-grained discourse relationships between sentences rather than mere similarity. As the first attempt, we do try to approximate the CST-relatedness between sentence pairs by various similarity measures.

Recently, (Marcu and Echihabi, 2002) presents an unsupervised approach to recognizing RST relationships that hold between spans of texts. Their method is inspiring but cannot be directly applied to our problem, for two reasons:

- CST relationships are more general and therefore more complicated than RST relations, due to their cross-document nature.
- Technically, (Marcu and Echihabi, 2002) still uses supervised learning techniques, but they can get large amount of training data by exploiting linguistic knowledge relatively easily. This doesn't hold in our experiment.

Another project related to ours is the METER corpus (Clough et al., 2002a), a resource for studying journalistic text reuse. METER contains sets of newswire and newspaper articles on the same topics. The texts have been manually annotated for various attributes, such as whether a given news article has been classified by an expert as being wholly-derived, partially-derived or not derived from the associated newswire. Recently, the corpus has been used in developing algorithms for automatically classifying the extent to which an input news article has been derived from the newswire (Clough et al., 2002b).

8. Conclusion

We described how we built the first CST-annotated news corpus, CST Bank. Currently, we are using it to work on the automatic identification of CST relationships in arbitrary clusters of related articles. We hope that our paper is a step forward towards better computational treatment of multiple, related texts.

9. References

Carletta, Jean, 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *CL*, 22(2):249–254.

Clough, Paul, Robert Gaizauskas, and S. L. Piao, 2002a. Building and Annotating a Corpus for the Study of Journalistic Text Reuse. In *Proceedings of LREC 2002*. Canary Islands, Spain.

Clough, Paul, Robert Gaizauskas, Scott S.L. Piao, and Yorick Wilks, 2002b. Measuring Text Reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hatzivassiloglou, Vasileios, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min-Yen Kan, and Kathleen R. McKeown, 2001. Simfinder: A flexible clustering tool for summarization. In *NAACL Workshop on Text Summarization*.

Mann, William and Sandra Thompson, 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text*, 8(3):243–281.

Marcu, Daniel and Abdessamad Echihabi, 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Radev, Dragomir R., 2000. A Common Theory of Information Fusion from Multiple Text Sources, Step One: Cross-document Structure. In *Proceedings of the 1st Workshop on Discourse and Dialogue of the Association for Computational Linguistics*. Hong Kong.

Radev, Dragomir R. and Kathleen R. McKeown, 1998. Generating Natural Language Summaries from Multiple On-Line Sources. *Computational Linguistics*, 4:469–500.