



6-2010

Creating Local Coherence: An Empirical Assessment

Annie Louis
University of Pennsylvania

Ani Nenkova
University of Pennsylvania, nenkova@cis.upenn.edu

Follow this and additional works at: http://repository.upenn.edu/cis_papers

 Part of the [Computer Sciences Commons](#)

Recommended Citation

Annie Louis and Ani Nenkova, "Creating Local Coherence: An Empirical Assessment", . June 2010.

Louis, A. & Nenkova, A., Creating Local Coherence: An Empirical Assessment, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, June 2010, doi: [anthology/N10-1043](https://doi.org/10.3115/anthology/N10-1043)

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/cis_papers/716
For more information, please contact libraryrepository@pobox.upenn.edu.

Creating Local Coherence: An Empirical Assessment

Abstract

Two of the mechanisms for creating natural transitions between adjacent sentences in a text, resulting in local coherence, involve discourse relations and switches of focus of attention between discourse entities. These two aspects of local coherence have been traditionally discussed and studied separately. But some empirical studies have given strong evidence for the necessity of understanding how the two types of coherence-creating devices interact. Here we present a joint corpus study of discourse relations and entity coherence exhibited in news texts from the Wall Street Journal and test several hypotheses expressed in earlier work about their interaction.

Disciplines

Computer Sciences

Comments

Louis, A. & Nenkova, A., Creating Local Coherence: An Empirical Assessment, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, June 2010, doi: [anthology/N10-1043](https://doi.org/10.3115/anthology/N10-1043)

Creating Local Coherence: An Empirical Assessment

Annie Louis

University of Pennsylvania
Philadelphia, PA 19104, USA
lannie@seas.upenn.edu

Ani Nenkova

University of Pennsylvania
Philadelphia, PA 19104, USA
nenkova@seas.upenn.edu

Abstract

Two of the mechanisms for creating natural transitions between adjacent sentences in a text, resulting in local coherence, involve discourse relations and switches of focus of attention between discourse entities. These two aspects of local coherence have been traditionally discussed and studied separately. But some empirical studies have given strong evidence for the necessity of understanding how the two types of coherence-creating devices interact. Here we present a joint corpus study of discourse relations and entity coherence exhibited in news texts from the Wall Street Journal and test several hypotheses expressed in earlier work about their interaction.

1 Introduction

Coherent discourse is characterized by local properties that are crucial for comprehension. In fact, a long line of linguistics and computational linguistics tradition has proposed that several levels of structure contribute to the creation of coherent discourse. Among these, the attentional structure (Grosz et al., 1995) and the relational structure (Mann and Thompson, 1988) of text, are the most widely discussed in the literature.

Centering theory (Grosz et al., 1995) is the dominant approach for describing and analyzing attentional structure. It assumes that readers of the text focus (center) their attention on a small number of salient discourse entities at a time and that there are preferred patterns for switching attention between entities mentioned in adjacent sentences. Relational

structure theories, on the other hand, describe how certain discourse (also called rhetorical or coherence) relations such as CONTRAST and CAUSE are inferred by the reader between adjacent units of text. The existence of richly annotated corpora and the development of automatic applications based on the theories have allowed empirical assessments of the validity and generality of these theories individually.

Such work has also provided increasingly strong evidence that attentional and relational structures need to be taken into account simultaneously. The motivation behind such proposals have been the empirical findings that “weak” discourse relations such as ELABORATION are the most common type of relations, and that a large percentage of adjacent sentences in fact do not have *any* entities in common.

In particular, a corpus based evaluation of Centering theory found that only 42% of pairs of adjacent sentences have at least one entity in common and hypothesized that discourse relations are responsible for creating local coherence in the remaining cases (Poesio et al., 2004). At the same time, the work of Knott et al. (2001) has discussed several theoretical complications arising from the existence of the very common and semantically weak ELABORATION relation. These researchers propose that replacing ELABORATION by an account of entity coherence such as Centering will be most beneficial. But work in information ordering (Karamanis, 2007) has not been able to confirm such claims that better results can be obtained by combining entity coherence with discourse relations.

Till recently, the absence of large corpora annotated for both discourse relations and coreference information has prohibited a detailed joint analysis of attentional and relational structure. We combine two

This work was partially supported by NSF Grant IIS -07-05671.

recently released corpora: discourse relations from the Penn Discourse Treebank and coreference annotations from the OntoNotes corpus, to assess the prevalence and interplay between factors that create local coherence in newspaper text.

Specifically, in our study we examine how three hypotheses formulated in prior work play out in the Wall Street Journal texts in our corpus:

Hypothesis 1 Adjacent sentences that do not share entities are related by non-elaboration discourse relations [Poesio et al. (2004) Sec. 5.2.2 Pg. 354].

Hypothesis 2 Adjacent sentences joined by non-elaboration discourse relation have lower entity coherence: such pairs are less likely to mention the same entities [Knott et al. (2001) Sec. 7 Pg. 10].

Hypothesis 3 Almost all pairs of sentences in a coherent text either share entities or participate in non-elaboration discourse relation (Knott et al., 2001; Poesio et al., 2004).

None of these hypotheses are validated. We find that only 38.65% of pairs that do not share entities participate in “core” relations such as temporal, contingency or comparison; the rate of coreference in these “core” relations is similar to that in weaker elaboration relations; about 30% of all sentence pairs neither share entities nor participate in a “core” discourse relation.

2 Data

In order to jointly analyze both discourse relations and noun phrase coreference between adjacent sentences, we combine annotations from two corpora, OntoNotes and the Penn Discourse Treebank. The two individual corpora are larger, but a smaller subset of 590 Wall Street Journal articles appear in both. All our analysis is for *adjacent sentences within paragraphs* in this subset of texts.

Penn Discourse Treebank The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) is the largest available corpus of annotations for discourse relations, covering one million words of the Wall Street Journal (WSJ). In the PDTB, two kinds of discourse relations are annotated. In *explicit relations*, a discourse connective such as “because”, “but” or “so” is present, as in the example below.

[Ex. 1] There is an incredible pressure on school systems and teachers to raise test scores. *So* efforts to beat the tests are

also on the rise.

On the other hand, relations can also exist without an explicit signal. In *Ex. 2*, it is clear that the second sentence is the *result* of the event mentioned in the first.

[Ex. 2] In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos. By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed.

Such relations are called *implicit relations*. In the PDTB, they are annotated between all adjacent sentences within a paragraph which do not already participate in an explicit discourse relation.

For both implicit and explicit relations, the semantic sense of the discourse relation is assigned from a hierarchy of senses. There are four classes of discourse relations at the topmost general level. The second level senses are shown within parentheses: Comparison (*Concession, Contrast, Pragmatic Concession/Contrast*), Contingency (*Cause, Condition, Pragmatic Cause/Condition*), Temporal (*Asynchronous, Synchronous*) and Expansion (*Alternative, Conjunction, Exception, Instantiation, List, Restatement*).

Some of the adjacent sentences in the texts, however, were found not to have a discourse relation between the events or propositions mentioned in them. Instead, they were related because both sentences mention the same entity, directly or indirectly, and the second sentence provides some further description of that entity. An *Entity Relation (EntRel)* was annotated for such sentence pairs as below.

[Ex. 3] Pierre Vincken, 61 years old, will join the board as a nonexecutive director Nov. 29. Mr. Vincken is chairman of Elsevier N.V., the Dutch publishing group.

OntoNotes For coreference information, we use the WSJ portion of the OntoNotes corpus version 2.9 (Hovy et al., 2006) which contains 590 documents. For these documents, we also have the PDTB annotations available. In OntoNotes, all noun phrases—pronouns, names and nominals are marked for coreference without any limitation to specific semantic categories.

3 Corpus study findings

For ease of presentation in the following analysis, we will call the Expansion and Entity relations “weak” and Temporal, Contingency and Compari-

0 to 100	0.41	500 to 1000	0.50
100 to 200	0.37	above 1000	0.51
200 to 500	0.48		

Table 1: Proportion of sentence pairs that *don't share* any entities for different document lengths (in words)

Type	Relation	No shared entities
Core	Temporal	122 (2.98)
	Contingency	752 (18.40)
	Comparison	706 (17.27)
Weak	Expansion	1870 (45.74)
	EntRel	638 (15.61)

Table 2: Distribution of the 4088 *non-entity sharing* sentence pairs. The proportions are shown in brackets.

son relations “core”, following the intuition that the semantics of the latter class is much more clearly defined. Implicit and explicit relations were not distinguished in the analysis.¹

Hypothesis 1 The first hypothesis is that adjacent sentences that do not share entities participate in core relations and so remain locally coherent.

Pairs of adjacent sentences that do not share any entities are common. In prior work, Poesio et al. (2004) found that 58% of adjacent sentence pairs in their corpus of museum object descriptions did not have overlapping mentions of any entity. The distribution in the WSJ texts is similar, as seen in Table 1. Depending on the length of the article, 37% to 51% of sentence pairs do not have any entity in common.²

Table 2 shows the distribution of discourse relations for the 4088 sentence pairs in the corpus that do not share any entities. Contrary to expectation, the majority of such pairs, 61%, are related by a weak discourse relation. Especially unexpected is the high percentage of *entity relations* (*EntRel*) that don't have actual entity overlap:

[Ex. 4] All four demonstrators were arrested. The law, which Bush allowed to take effect without his signature, went into force Friday.

[Ex. 5] Authorities in Hawaii said the wreckage of a missing commuter plane with 20 people aboard was spotted in a remote valley on the island of Molokai. There wasn't any evidence of survivors.

¹For brevity we present combined results for both implicit and explicit relations. However, most of our conclusions remain the same when the two types are distinguished.

²There are around 100 documents in each length range.

Type	Relation	Total	Share entities
Core	Temporal	365	243 (66.57)
	Contingency	1570	818 (52.10)
	Comparison	1477	771 (52.20)
Weak	Expansion	3569	1699 (47.60)
	EntRel	1424	786 (55.20)

Table 3: Rate of *entity sharing*

	Share entities	No sharing
core relations	1832 (21.80)	1580 (18.80)
weak relations	2485 (29.56)	2508 (29.84)

Table 4: Total number (proportion) of sentence pairs in the corpus in the given categories

Hypothesis 2 The second hypothesis states that adjacent sentences joined by a core discourse relation are less likely to mention the same entities in comparison to weak relations. But as Table 3 reveals, this is generally not the case.

Adjacent sentences in Temporal relation are very likely to share entities—almost 70% of them do. Over half of all Contingency and Comparison relations also share entities. But, the rates of sharing in Comparison and Contingency relations are significantly lower compared to Entity relations (under a two-sided binomial test). *Ex. 6* shows a Contingency relation without entity sharing.

[Ex. 6] Without machines, good farms can't get bigger. So the potato crop, once 47 million tons, is down to 35.

However, adjacent sentences with Expansion relation turn out least likely to share entities. The entity sharing rates of all the other relations were found to be significantly higher than Expansion.

Hypothesis 3 Finally, we test the hypothesis that the majority of adjacent sentences exhibit coherence because they either share entities or form the arguments of a core discourse relation.

This hypothesis is not supported in the WSJ data (see Table 4): 30% of all sentence pairs are in a weak discourse relation—Expansion or EntRel—and do not have any entities in common. In a sense, neither of the theories of entity or relational coherence can explain what mechanisms create the local coherence for these pairs. In order to glean some insights of how coherence is created there, we examine the behavior of different types of Elaboration relations (Table 5). There is a wide variation between the dif-

Alternative	67 (0.63)	Instantiation	490 (0.34)
Restatement	960 (0.56)	List	165 (0.28)
Conjunction	1021 (0.48)	EntRel	1424 (0.55)

Table 5: Total number of different Expansion relations and their coreference probability (in brackets)

ferent types of Expansions.

When the function of an expansion sentence is to provide an *alternative explanation* or *restate* an utterance, the probability of entity overlap is very high and patterns similarly with entity relations (around 60%). Below is an example restatement sentence.

[Ex. 7] {Researchers in Belgium}_r said {they}_r have developed a genetic engineering technique for creating hybrid plants for a number of crops, such as cotton, soybeans and rice. {The scientists at Plant Genetic Systems N.V.}_r isolated a gene that could lead to a generation of plants possessing a high-production trait.

However, the two classes—Instantiation and List largely appear with very little entity overlap, 37% and 29% respectively. An Instantiation relation is used to provide an example and hence has little overlap with the previous sentence (Ex. 8 and 9).

[Ex. 8] There may be a few cases where the law breaking is well pinpointed and so completely non-invasive of the rights of others that it is difficult to criticize it. The case of Rosa Parks, the black woman who refused to sit at the back of the bus, comes to mind as an illustration.

[Ex. 9] The economy is showing signs of weakness, particularly among manufacturers. Exports, which played a key role in fueling growth over the last two years, seem to have stalled.

List relations, on the other hand, connect sentences where each of them elaborates on a common proposition mentioned earlier in the discourse. Here is an example five sentence paragraph with list relations but no entity repetition at all.

[Ex. 10] Designs call for a L-shaped structure with a playground in the center. [Implicit List] There will be recreation and movie rooms. [Implicit List] Teens will be able to listen to music with headsets. [Implicit List] Study halls, complete with reference materials will be available. [Explicit List] And there will be a nurse’s station and rooms for children to meet with the social workers.

In Ex. 10, as well as some others, a broad notion of entity coherence—bridging anaphora can be applied. Poesio et al. (2004) also note this fact, but

also say that such instances are very difficult to annotate reliably. Our work is based on coreference annotations which can be marked with considerably high inter-annotator agreement.

4 Conclusions

The recent release of corpora annotated for coreference and discourse relations for the same texts have made possible to empirically assess claims about the interaction between two types of local coherence: relational and entity. We find that about half of the pairs of adjacent sentences do not share any entities at all, and about 60% are related by weak discourse relations. We test the hypothesis from prior work that these two types of coherence are complementary to each other and taken together explain most local coherence. We find that the two types of coherence mechanisms are neither mutually exclusive nor do they explain all the data. Future work in discourse analysis will need to develop better understanding of how the two types of coherence interact.

References

- B. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: a framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–226.
- E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of NAACL-HLT*, pages 57–60.
- N. Karamanis. 2007. Supplementing entity coherence with local rhetorical relations for information ordering. *Journal of Logic, Language and Information*, 16(4):445–464.
- A. Knott, J. Oberlander, M. O’Donnell, and C. Mellish. 2001. Beyond elaboration: the interaction of relations and focus in coherent text. In *Text Representation: Linguistic and Psycholinguistic Aspects, chapter 7*, pages 181–196.
- W.C. Mann and S.A. Thompson. 1988. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8.
- M. Poesio, R. Stevenson, B. Di Eugenio, and J. Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, pages 309–363.
- R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of LREC’08*.