Constructing a Textual Semantic Relation Corpus Using a Discourse Treebank

Rui Wang and Caroline Sporleder

Computational Linguistics
Saarland University
{rwang,csporled}@coli.uni-sb.de

Abstract

In this paper, we present our work on constructing a textual semantic relation corpus by making use of an existing treebank annotated with discourse relations. We extract adjacent text span pairs and group them into six categories according to the different discourse relations between them. After that, we present the details of our annotation scheme, which includes six textual semantic relations, *backward entailment*, *forward entailment*, *equality*, *contradiction*, *overlapping*, and *independent*. We also discuss some ambiguous examples to show the difficulty of such annotation task, which cannot be easily done by an automatic mapping between discourse relations and semantic relations. We have two annotators and each of them performs the task twice. The basic statistics on the constructed corpus looks promising: we achieve 81.17% of agreement on the six semantic relation annotation with a .718 kappa score, and it increases to 91.21% if we collapse the last two labels with a .775 kappa score.

1. Introduction

Traditionally, the term 'semantic relation' refers to relations that hold between the meanings of two words, e.g., synonymy, hypernymy, etc. These relations are usually situation-independent. Discourse relations, on the other hand, tend to depend on the wider context of an utterance. However, the term 'semantic relation' has also been used in a wider sense to refer to relations between two linguistic expressions or texts, such as paraphrasing, textual entailment, etc. (Murakami et al., 2009). We call such relations *Textual Semantic Relations (TSRs)*.

In this paper, we investigate the connection between discourse relations, such as CAUSE or CONTRAST, and TSRs, such as ENTAILMENT or CONTRADICTION. In general, (strict) entailment or repetition is unlikely to appear frequently in a naturally occurring discourse since redundant information content would violate the Gricean maxim of Manner (Grice, 1975). Nonetheless, there are situations in which information is at least partly repeated, e.g., in restatements or summaries.

So far, there has been little work that has investigated the connection between discourse and semantic relations. Most research on textual inference focuses on the lexical, syntactic, and semantic levels. Furthermore, studies that have looked at the discourse level have typically been restricted to specific discourse context, for example, on whether examples of entailment can be acquired from news texts and their corresponding headlines (Burger and Ferro, 2005).

There has been some previous research on constructing corpora on TSRs. One large collection is provided by the *Recognizing Textual Entailment* (RTE) community, following each year's challenge, from RTE-1 in 2005 (Dagan et al., 2005) till now (Bentivogli et al., 2009). The corpora from the first two RTE challenges were annotated with two labels: One is YES, meaning that there is an entailment relation from the first text, *Text* (**T**), to the second text, *Hypothesis* (**H**); and the other label is NO, meaning there is no such relation. Starting from the RTE-3 Pilot task,[1] the annotation is extended to three labels, ENTAILMENT, CONTRADICTION, and UNKNOWN. ENTAILMENT is the same as the previous YES; but NO is divided into CONTRADICTION and UNKNOWN, to differentiate cases where **T** and **H** are contradictory to each other from all the other cases. The RTE data are acquired from other natural language processing tasks, like information retrieval, question answering, summarization, etc., and thus, in some sense, the corpora are more application-driven than linguistically motivated.

The FraCas dataset (Cooper et al., 1996) focuses more on the linguistic side, aiming to cover different linguistic/semantic phenomena. The annotation is also three-way. However, this dataset is manually constructed, sentences are carefully selected, and it turns out to have a "text-book" style, which is quite different from the real data we usually need to process. MacCartney and Manning (2007) have studied on some possible TSRs between texts. In addition to the three relations used by the RTE community, they propose two extra relations, INDEPENDENT and EXCLUSIVE, and also separate ENTAILMENT relation into two, FORWARD ENTAILMENT and BACKWARD ENTAILMENT, according to the direction. Some other researchers (Wang and Zhang, 2009) also suggest a relatedness measurement, which covers ENTAILMENT and CONTRADICTION, and their proposal is based on empirical results. All these previous studies enlighten us to further explore the semantic relations between two texts, not to mention a large collection of papers on paraphrase (which can be viewed as a bi-directional entailment relation) acquisition and application in the literature (Shinyama et al., 2002; Barzilay and Lee, 2003, etc.).

In this work, we analyze the relationship between discourse relations and semantic relations, based on existing resources and our observations. By doing that, we aim to achieve:

1. a better understanding of both relations (e.g., whether specific TSRs tend to co-occur with specific discourse relations)

2. building a useful corpus on textual semantic relations for future applications, e.g. the RTE task.

[1] http://nlp.stanford.edu/RTE3-pilot/

2. Annotation

To obtain data annotated with discourse relations, we used the RST Discourse Treebank (RST-DT).[2] RST defines a set of 24-30 relatively fine-grained discourse relations, such as CONTRAST, RESTATEMENT, ELABORATION or BACKGROUND. Most relations are binary and link a *nucleus* (N) (i.e., a more important text span) to a *satellite* (S) (i.e., the less important span). We extracted all relations holding between adjacent sentences from the RST-DT, thus excluding relations between sub-sentential clauses or larger pieces of text.

By looking at the discourse relations mentioned above, we can already observe some potentially relevant TSRs. For instance, if two adjacent texts have the RESTATEMENT relation, they could be a (non-strict) paraphrase to each other. An ELABORATION relation can exist between two texts, where a backward entailment might also hold, e.g., a concrete story entails a short headline. A CONTRAST relation may contain a contradiction between two texts, although people usually do not make totally contradictory utterances. In the most common situation, when the two text spans have no such strong TSRs (e.g. the BACKGROUND relation), we assume that they are still relevant to each other in some sense. They may mention the same entity, different steps of one procedure, consequent actions, etc.

Since the inventory of RST relations is relatively fine-grained, we manually grouped the relations into six classes, more or less following the "relation groups" in the RST-DT annotation manual.[3] Each group contains related discourse relations and we hypothesize that relations within a given group also behave similar with respect to the TSRs to which they can be mapped. The resulting six relation groups are:

- background : BACKGROUND, CIRCUMSTANCE;
- elaboration : ELABORATION-SET-MEMBER, ELABORATION-PROCESS-STEP, ELABORATION-OBJECT-ATTRIBUTE, ELABORATION-GENERAL-SPECIFIC, EXAMPLE, ELABORATION-ADDITIONAL;
- explanation : EXPLANATION-ARGUMENTATIVE, EVIDENCE, PURPOSE, REASON;
- consequence : CONSEQUENCEN, CONSEQUENCES, CONSEQUENCE, CAUSE, CAUSE-RESULT, RESULT;
- contrast : ANTITHESIS, CONCESSION, CONTRAST, INTERPRETATIONS, INTERPRETATIONN;
- restatement : RESTATEMENT, SUMMARYS, SUMMARYN.

We excluded ENABLEMENT, which is grouped with PURPOSE in the RST-DT manual, because the nucleus in ENABLEMENT is supposed to be unrealized. We also excluded EVALUATION, which is grouped with INTERPRETATION, but unlike INTERPRETATION, both text spans of EVALUATION are "attributed to the same agent", i.e. there is no contrastive aspect. The rest of the excluded relations, e.g., LIST, SEQUENCE, etc., were disregarded due to one of two reasons: 1) we hypothesize that these relations are not interesting for semantic relations, especially for the entailment relation; and 2) some of them occur very infrequently in the corpus, making it impossible to make any empirical statements about them.

The extracted RST-DT examples were then manually labelled with TSRs. We define eight annotation labels:

- FE - Forward Entailment : There is an entailment relation between the two text spans, and the direction is from the first one to the second one.
- BE - Backward Entailment : There is an entailment relation between the two spans, and the direction is from the second one to the first one, e.g. Example 4 in Table 2.
- E - Equality : The two spans are paraphrases of each other, or the entailment relation holds in both directions (forward and backward). The meaning is (almost) the same, like Example 12 in Table 2.
- C - Contradiction : There is a contradiction between the two spans. The meaning or information of (some parts of) the two spans are contradictory to each other. For instance, Example 6 in Table 2.
- O - Overlapping : None of the above relations holds, but the spans are relevant to each other and share much meaning or information, like Example 1 in Table 2.
- I - Independent : There are no overlapping events between two text spans, even though there might be one entity mentioned in both, like Example 11 in Table 2.
- ? - Uncertain : The question mark can be combined with the first four labels, meaning that the relation holds not strictly, but loosely from the annotator's point of view. For instance, Example 5 in Table 2 is not a strict FE, but the information in the second span can be inferred from the first span with a relatively high probability.
- F - False : The example is not valid. It could be that the sentence extracted from the corpus is incomplete or hard to understand without further context.

Our goal was to capture the whole spectrum of different relations between meanings of two texts. On the dimension of overlapping information, we have little overlapping information (i.e. I), some overlapping (i.e. O), and fully the same (i.e. E); on the dimension of consistency, we have both contradictory relation (i.e. C) and consistent relations (i.e. all the other relations). In particular, we also incorporate the directionality of the ENTAILMENT relation (i.e. FE and BE), which has not been fully explored in the field yet.

[2] Available from the LDC: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T07

[3] http://www.isi.edu/~marcu/software/manual.ps.gz

The annotations were done by two annotators. Annotating TSRs is a relatively hard task, particularly when it is done on naturally occurring examples because, as was mentioned before, totally clear cases of entailment and contradiction are relatively rare compared to artificially constructed examples. To arrive at a reasonably reliable annotation, the annotation was done in two rounds. Initially, the annotators only labelled a subset of the data (100 examples). The annotators then discussed examples on which they disagreed with the aim of arriving at a more consistent annotation. The discussion phase also helped in making the annotation guidelines more precise. In the second round, the remaining examples were labelled. So far, we have annotated 319 text pairs, and among them there are 239 (75%) valid pairs, i.e., not labelled as F.

To assess the reliability of our annotation, we computed the inter-annotator agreement (excluding instances labelled as F). The results are shown in Table 1. Under the 'strict' agreement evaluation scheme labels with and without a question mark (e.g., FE vs. FE?) were considered different, under a 'loose' evaluation scheme the question marks were disregarded. We also computed the agreement after collapsing the classes 'independent' (I) and 'overlap' (O), since these two classes proved often difficult to distinguish (see Section 3.) and moreover their distinction is less relevant for our present study, which focuses on entailment and contradiction. The inter-annotator agreement for our data lies between 79% and 91% and is thus relatively high. We also computed the Kappa statistic (Krippendorff, 1980), which corrects the percentage agreement for expected chance agreement. Our kappa scores range from .696 to .775 (see Table 1), which is considered as good agreement. Hence our annotations are generally fairly reliable.

In the next section, we will provide some example annotations to make the definition of the TSRs more concrete and we will also discuss some borderline cases.

3. Example Annotations and Difficult Cases

To illustrate our annotation scheme, we show some examples from our data in Table 2. Generally our annotators agreed well on which TSR to assign to a given example (see Section 2.). However, some distinctions proved difficult to make. In this section, we discuss examples, for which the distinction between two labels is not straightforward. An annotation decision that proved particularly difficult was the distinction between I and O. In practice, we use the number of shared entities as one criteria, namely, I allows at most one shared entity between the two text spans, while examples with a higher overlap should be labelled O (unless one of the other relations holds). A relatively clear case of I is Example 11 in Table 2, where there are no obvious shared entities between the two spans. In contrast, Example 1 is a fairly straightforward case of an overlap relation (O): "The engineering company" is co-referent with "Larsen & Toubro" and "the giant Reliance textile group" is co-referent with "its new owners".

Although the distinction is easy for most of the cases, there are still some tricky ones. For instance, in Example 7, both annotators agree that both spans evoke a reference to "sales" but one annotator is not sure whether "some traders" in the first span are the same as "platinum and palladium traders" in the second span. Example 10 is more interesting. "These goals" in the second span are generally referring to those mentioned in the proposal (from the first span), but it is unclear whether they should be treated as single entity or multiple entities.

Example 5 illustrates the use of a question mark in combination with one of the annotation labels. In this examples it is difficult to verify the quantifier "every" in the first text, but we still think the forward entailment relation holds, albeit loosely. As we mentioned at the beginning, it is almost impossible to find strict entailment or repetition in a naturally occurring discourse since it violates the Gricean maxim of Manner. Instead one finds cases of 'soft entailment' where one span follows from the other with a reasonably high probability. Annotators sometimes differ with respect to how high they estimate this probability to be, and annotate FE or FE?, depending on their own interpretation of the likelihood of entailment.

Entailment relations might also be interpreted differently. The annotators agreed on the BE relation for Example 4 in Table 2, while Example 2 and Example 8 are not agreed on. In Example 2, one annotator considers that a big project does not necessarily mean the contract signing is important (i.e. I), while the other annotator understands "big" as "financially significant", which does entail "important" (i.e. BE). In Example 8, one annotator thinks "the selling of cyclical stocks" does not entail "recession fears" (i.e. I), while the other annotator feels that "sell-off" gives an impression of such fears (i.e. BE). In addition, these examples containing abstraction and inference could hardly be labeled as O, since shared (concrete) entities are difficult to find.

For contradiction cases, both annotators agree on Example 6 in Table 2, since there is a sharp contrast between what "Wang executives had said" in the summer and what they (i.e. "Mr. Miller") say now. However, they disagreement on Example 9. One annotator interprets "in May" as an implicit mentioning of (May of) "this year", which is contradictory to "more than five years ago" in the other text; while the other annotator does not consider them comparable to each other, thus annotating O.

4. Corpus Statistics

The annotation still needs to be finalized, therefore, we will present some preliminary results on the current version of the corpus we are constructing.[4]

For this study, we were particularly interested in whether specific discourse relations tend to correlate with particular TSRs. Table 3 provides some basic statistics of the corpus, as well as the distribution of the discourse relation groups versus the TSR annotations. Note that we only count the agreed text pairs in this table.

It can be seen that I and O are the most frequent relations, holding between 50.52% and 28.84% of the text pairs, re-

[4]Due to the fact that the the RST Discourse Treebank is licensed, we cannot release the annotated corpus directly. Please contact the authors if you are interested in the corpus.

	strict		loose	
Annotations	Six-Way	Collapse I&O	Six-Way	Collapse I&O
Agreement	79.50%	89.54%	81.17%	91.21%
Kappa	.696	.736	.718	.775

Table 1: Inter-Annotator Agreement. The difference between *strict* and *loose* is that the latter ignores the question mark in the annotations. And "Collapse I&O" means we treat I and O as one relation.

Id	Relation Groups	Sentences	Anno_1	Anno_2
1	background	**The engineering company** was acquired in a takeover earlier this year by **the giant Reliance textile group**.	O	O
		Although **Larsen & Toubro** hadn't raised money from the public in 38 years, **its new owners** frequently raise funds on the local market.		
2	elaboration	The contract signing represented a major step in the long-planned petrochemical project.	I	BE
		At an estimated $360 million, the project would represent the single largest foreign investment in the Philippines since President Corazon Aquino took office in February 1986.		
3	elaboration	**Eli Lilly & Co.**, the Indianapolis-based drug manufacturer, dominates the U.S. human insulin market with its product known as Humulin.	I	O
		Lilly is building plants to make the insulin in Indianapolis and Fagershein, France.		
4	explanation	Ford Motor Co. and Chrysler Corp. representatives criticized Mr. Tonkin's plan as unworkable.	BE	BE
		It "is going to sound neat to the dealer except when his 15-day car supply doesn't include the bright red one that the lady wants to buy and she goes up the street to buy one," a Chrysler spokesman said.		
5	explanation	Many of the problems you presented exist in every part of this country.	FE?	FE?
		Poverty is only two blocks from President Bush's residence.		
6	explanation	In response to questions after the annual meeting, Mr. Miller said the company is no longer looking for an equity investor.	C	C
		During the summer, Wang executives had said they might seek outside investment.		
7	explanation	Some traders were thought to be waiting for **the auto sales** report, which will be released today.	O	I
		Such sales are watched closely by platinum and palladium traders because both metals are used in automobile catalytic converters.		
8	consequence	Recession fears are springing up again among investors.	BE	I
		Analysts say that the selling of cyclical stocks yesterday will be followed by a sell-off in shares of companies with big debt loads on their balance sheets.		
9	contrast	Gulf Power said in May that an internal audit had disclosed that at least one vendor had used false invoices to fund political causes.	C	O
		But the company said the political contributions had been made more than five years ago.		
10	contrast	The proposal reiterates the U.S. desire to scrap or reduce a host of trade-distorting subsidies on farm products.	I	O
		But it would allow considerable flexibility in determining how and when **these goals** would be achieved.		
11	contrast	Rates are determined by the difference between the purchase price and face value.	I	I
		Thus, higher bidding narrows the investor's return while lower bidding widens it.		
12	restatement	"Anne doesn't believe in blandness," said Ms. Smith.	E	E
		"She wants things to be exciting."		

Table 2: Examples of the annotated text pairs for different discourse relations

	I	O	C	FE	BE	E	all	%
BACKGROUND	18	12	0	0	0	0	30	15.46%
CONSEQUENCE	8	6	0	3	1	0	18	9.28%
CONTRAST	22	11	13	1	2	0	49	25.26%
ELABORATION	29	17	1	0	4	1	52	26.80%
EXPLANATION	21	7	1	1	9	0	39	20.10%
RESTATEMENT	0	1	0	1	2	2	6	3.09%
all	98	54	15	6	18	3	194	100.0%
%	50.52%	28.84%	7.73%	3.09%	9.28%	1.55%	100.0%	

Table 3: Distribution of the Annotation Labels across the Relation Groups

spectively. The other relations are comparably rare, especially true, bi-directional entailment (E), which only occurs three times. This is not surprising since we hypothesized that true entailment would be rare in naturally occurring text. Backward entailment BE is more frequent than forward entailment FE, and contradictions (C) are more or less equally frequent as backward entailments.

With respect to discourse relations, CONTRAST, ELABORATION, and EXPLANATION occur most often in our sample and these three relations are more or less equally frequent. While our sample is a bit biased with respect to discourse relations, since we excluded some relations, the fact that these three relations are relatively frequent between adjacent text spans is to be expected. RESTATEMENT is the least frequent relation, which is also expected.

Given the relatively small data set, it is difficult to make definite statements about the correlation of different discourse relations and TSRs, however, some trends are observable. First, it seems that TSRs distribute unevenly across different discourse relations. For instance, CONTRAST contains almost all the C cases (13/15), while ELABORATION and EXPLANATION have the most BE cases (4/18 and 9/18). As expected, RESTATEMENT relations tend to correlate with some form of entailment (E, BE, or FE), five out of six restatements involve entailment.

It is also interesting to look at the unobserved pairings of discourse relation and TSR. Some of these seem very plausible. For instance, one would not expect contradiction or independence for a RESTATEMENT relation. Likewise, one would not expect to find a bi-directional entailment for a CONTRAST relation.

However, while some trends are observable and intuitive, it is also clear from the data that there is no clear one-to-many or many-to-one mapping between discourse relations and TSRs. Most discourse relations can co-occur with most TSRs and vice versa. This suggests that discourse relations and TSRs capture different and partly independent aspects of meaning.

5. Conclusion

In this paper, we presented our work on constructing a textual semantic relation corpus on top of a treebank annotated with discourse relations. We extracted text span pairs related by different discourse relations (from six broad relation groups) and annotated each pair with one of six semantic relations. Despite the fact that it is difficult to find totally clear-cut examples of semantic relations such as entailment or contradiction in naturally occurring examples of adjacent text spans, we do obtain a relatively high inter-annotator agreement.

An initial analysis of the annotated data revealed some interesting trends for correlations between discourse relations and semantic relations. However, to draw reliable empirical conclusions a larger data set is required. To this end, we plan to annotate more data. This will allow us to perform a more detailed analysis. For instance, we would like to investigate whether all discourse relations within a group behave similarly or whether individual discourse relations correlate more or less with particular semantic relations. We also plan to look at the influence of *nucleus* and *satellite*. Furthermore, in order to fully evaluate the corpus, we will test an existing RTE system on it to see how "difficult" it is to label automatically, compared with other datasets.

Acknowledgements

The first author is supported by the PIRE scholarship program and this work was partially funded by the German Research Foundation DFG as part of the MMCI Cluster of Excellence.

6. References

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 16–23, Edmonton, Canada, May 27-June 01. Association for Computational Linguistics.

L. Bentivogli, B. Magnini, I. Dagan, H.T. Dang, and D. Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *Proceedings of the Text Analysis Conference (TAC 2009) Workshop*, Gaithersburg, Maryland, USA, November. National Institute of Standards and Technology.

John Burger and Lisa Ferro. 2005. Generating an entailment corpus from news headlines. In *Proceedings of the ACL Workshop om Empirical Modelling of Semantic Equivalence and Entailment*, pages 49–54.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. A framework for computational semantics (FraCaS). Technical report, The FraCaS Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In Quiñonero-Candela et al., editor, *MLCW 2005*, volume LNAI Volume 3944, pages 177–190. Springer-Verlag.

H. P. Grice. 1975. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Synax and Semantics*, volume 3 (Speech Acts), pages 41–58. Academic Press.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, pages 193–200, Prague, Czech Republic. Association for Computational Linguistics.

Koji Murakami, Shouko Masuda, Suguru Matsuyoshi, Eric Nichols, Kentaro Inui, and Yuji Matsumoto. 2009. Annotating semantic relations combining facts and opinions. In *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP-09*, pages 150–153.

Yusuke Shinyama, Satoshi Sekine, Kiyoshi Sudo, and Ralph Grishman. 2002. Automatic paraphrase acquisition from news articles. In *Proceedings of Human Language Technology Conference (HLT 2002)*, San Diego, USA. Association for Computational Linguistics.

Rui Wang and Yi Zhang. 2009. Recognizing textual relatedness with predicate-argument structures. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, Singapore, Singapore. Association for Computational Linguistics.