Towards the Unsupervised Acquisition of Discourse Relations

Christian Chiarcos
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292
chiarcos@daad-alumni.de

Abstract

This paper describes a novel approach towards the empirical approximation of discourse relations between different utterances in texts. Following the idea that every pair of events comes with preferences regarding the range and frequency of discourse relations connecting both parts, the paper investigates whether these preferences are manifested in the distribution of relation words (that serve to signal these relations).

Experiments on two large-scale English web corpora show that significant correlations between pairs of adjacent events and relation words exist, that they are reproducible on different data sets, and for three relation words, that their distribution corresponds to theory-based assumptions.

1 Motivation

Texts are not merely accumulations of isolated utterances, but the arrangement of utterances conveys *meaning*; human text understanding can thus be described as a process to recover the global structure of texts and the relations linking its different parts (Vallduví 1992; Gernsbacher et al. 2004). To capture these aspects of meaning in NLP, it is necessary to develop operationalizable theories, and, within a supervised approach, large amounts of annotated training data. To facilitate manual annotation, weakly supervised or unsupervised techniques can be applied as preprocessing step for *semi*manual annotation, and this is part of the motivation of the approach described here.

Discourse relations involve different aspects of meaning. This may include factual knowledge about the connected discourse segments (a 'subject-matter' relation, e.g., if one utterance represents the cause for another, Mann and Thompson 1988, p.257), argumentative purposes (a 'presentational' relation, e.g., one utterance motivates the reader to accept a claim formulated in another utterance, ibid., p.257), or relations between entities mentioned in the connected discourse segments (anaphoric relations, Webber et al. 2003). Discourse relations can be indicated explicitly by optional cues, e.g., adverbials (e.g., *however*), conjunctions (e.g., *but*), or complex phrases (e.g., *in contrast to what Peter said a minute ago*). Here, these cues are referred to as *relation words*.

Assuming that relation words are associated with specific discourse relations (Knott and Dale 1994; Prasad et al. 2008), the distribution of relation words found between two (types of) events can yield insights into the range of discourse relations possible at this occasion and their respective likeliness. For this purpose, this paper proposes a background knowledge base (BKB) that hosts pairs of events (here heuristically represented by verbs) along with distributional profiles for relation words. The primary data structure of the BKB is a *triple* where one event (type) is connected with a particular relation word to another event (type). Triples are further augmented with a *frequency score* (expressing the likelihood of the triple to be observed), a *significance score* (see below), and a *correlation score* (indicating whether a pair of events has a positive or negative correlation with a particular relation word).

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 213–217,
Jeju, Republic of Korea, 8-14 July 2012.

Triples can be easily acquired from automatically parsed corpora. While the relation word is usually part of the utterance that represents the source of the relation, determining the appropriate target (antecedent) of the relation may be difficult to achieve. As a heuristic, an adjacency preference is adopted, i.e., the target is identified with the main event of the preceding utterance.[1] The BKB can be constructed from a sufficiently large corpus as follows:

- identify event types and relation words
- for every utterance
  - create a candidate triple consisting of the event type of the utterance, the relation word, and the event type of the preceding utterance.
  - add the candidate triple to the BKB, if it found in the BKB, increase its score by (or initialize it with) 1,
- perform a pruning on all candidate triples, calculate significance and correlation scores

Pruning uses statistical significance tests to evaluate whether the relative frequency of a relation word for a pair of events is significantly higher or lower than the relative frequency of the relation word in the entire corpus. Assuming that incorrect candidate triples (i.e., where the factual target of the relation was non-adjacent) are equally distributed, they should be filtered out by the significance tests.

The goal of this paper is to evaluate the validity of this approach.

2 Experimental Setup

By generalizing over multiple occurrences of the same events (or, more precisely, event types), one can identify preferences of event pairs for one or several relation words. These preferences capture *context-invariant* characteristics of pairs of events and are thus to considered to reflect a semantic predisposition for a particular discourse relation.

Formally, an event is the semantic representation of the meaning conveyed in the utterance. We assume that the same event can reoccur in different contexts, we are thus studying relations between *types* of events. For the experiment described here, events are heuristically identified with the main predicates of a sentence, i.e., non-auxiliar, non-causative, non-modal verbal lexemes that serve as heads of main clauses.

The primary data structure of the approach described here is a triple consisting of a source event, a relation word and a target (antecedent) event. These triples are harvested from large syntactically annotated corpora. For intersentential relations, the target is identified with the event of the immediately preceding main clause. These extraction preferences are heuristic approximations, and thus, an additional pruning step is necessary.

For this purpose, statistical significance tests are adopted (χ^2 for triples of frequent events and relation words, t-test for rare events and/or relation words) that compare the relative frequency of a relation word given a pair of events with the relative frequency of the relation word in the entire corpus. All results with $p \geq .05$ are excluded, i.e., only triples are preserved for which the observed positive or negative correlation between a pair of events and a relation word is not due to chance with at least 95% probability. Assuming an even distribution of incorrect target events, this should rule these out. Additionally, it also serves as a means of evaluation. Using statistical significance tests as pruning criterion entails that all triples eventually confirmed are statistically significant.[2]

This setup requires *immense* amounts of data: We are dealing with several thousand events (theoretically, the total number of verbs of a language). The chance probability for two events to occur in adjacent position is thus far below 10^{-6}, and it decreases further if the likelihood of a relation word is taken into consideration. All things being equal, we thus need *millions* of sentences to create the BKB.

Here, two large-scale corpora of English are employed, PukWaC and Wackypedia_EN (Baroni et al. 2009). PukWaC is a 2G-token web corpus of British English crawled from the `uk` domain (Ferraresi et al.

[1] Relations between non-adjacent utterances are constrained by the structure of discourse (Webber 1991), and thus less likely than relations between adjacent utterances.

[2] Subsequent studies may employ less rigid pruning criteria. For the purpose of the current paper, however, the statistical significance of all extracted triples serves as an criterion to evaluate methodological validity.

2008), and parsed with MaltParser (Nivre et al. 2006). It is distributed in 5 parts; Only PukWaC-1 to PukWaC-4 were considered here, constituting 82.2% (72.5M sentences) of the entire corpus, PukWaC-5 is left untouched for forthcoming evaluation experiments. Wackypedia_EN is a 0.8G-token dump of the English Wikipedia, annotated with the same tools. It is distributed in 4 different files; the last portion was left untouched for forthcoming evaluation experiments. The portion analyzed here comprises 33.2M sentences, 75.9% of the corpus.

The extraction of events in these corpora uses simple patterns that combine dependency information and part-of-speech tags to retrieve the main verbs and store their lemmata as event types. The target (antecedent) event was identified with the last main event of the preceding sentence. As relation words, only sentence-initial children of the source event that were annotated as adverbial modifiers, verb modifiers or conjunctions were considered.

3 Evaluation

To evaluate the validity of the approach, three fundamental questions need to be addressed: **significance** (are there significant correlations between pairs of events and relation words ?), **reproducibility** (can these correlations confirmed on independent data sets ?), and **interpretability** (can these correlations be interpreted in terms of theoretically-defined discourse relations ?).

3.1 Significance and Reproducibility

Significance tests are part of the pruning stage of the algorithm. Therefore, the number of triples eventually retrieved confirms the existence of statistically significant correlations between pairs of events and relation words. The left column of Tab. 1 shows the number of triples obtained from PukWaC subcorpora of different size.

For reproducibility, compare the triples identified with Wackypedia_EN and PukWaC subcorpora of different size: Table 1 shows the number of triples found in both Wackypedia_EN and PukWaC, and the *agreement* between both resources. For two triples involving the same events (event types) and the same relation word, agreement means that the relation word shows either positive or negative correlation in both corpora, disagreement means positive correlation in one corpus and negative correlation in the other.

PukWaC (sub)corpus		Wackypedia_EN triples		
sentences	triples	common	agreeing	%
1.2M	74	20	12	60.0
4.8M	832	177	132	75.5
19.2M	7,342	938	809	86.3
38.4M	20,106	1,783	1,596	89.9
72.5M	46,680	2,643	2,393	90.5

Table 1: Agreement with respect to positive or negative correlation of event pairs and relation words between Wackypedia_EN and PukWaC subcorpora of different size

	PukWaC triples			agreement (%)	
	total	vs. H	vs. T	vs. H	vs. T
B: *but*	11,042	6,805	1,525	97.7	62.2
H: *however*	7,251		1,413		66.9
T: *then*	1,791				

Table 2: Agreement between *but* (B), *however* (H) and *then* (T) on PukWaC

Table 1 confirms that results obtained on one resource can be reproduced on another. This indicates that triples indeed capture context-invariant, and hence, semantic, characteristics of the relation between events. The data also indicates that reproducibility increases with the size of corpora from which a BKB is built.

3.2 Interpretability

Any theory of discourse relations would predict that relation words with similar function should have similar distributions, whereas one would expect different distributions for functionally unrelated relation words. These expectations are tested here for three of the most frequent relation words found in the corpora, i.e., *but*, *then* and *however*. *But* and *however* can be grouped together under a generalized notion of contrast (Knott and Dale 1994; Prasad et al. 2008); *then*, on the other hand, indicates a temporal and/or causal relation.

Table 2 confirms the expectation that event pairs that are correlated with *but* tend to show the same correlation with *however*, but not with *then*.

4 Discussion and Outlook

This paper described a novel approach towards the unsupervised acquisition of discourse relations, with encouraging preliminary results: Large collections of parsed text are used to assess distributional profiles of relation words that indicate discourse relations that are possible between specific types of events; on this basis, a background knowledge base (BKB) was created that can be used to predict an appropriate discourse marker to connect two utterances with no overt relation word.

This information can be used, for example, to facilitate the semiautomated annotation of discourse relations, by pointing out the 'default' relation word for a given pair of events. Similarly, Zhou et al. (2010) used a language model to predict discourse markers for implicitly realized discourse relations. As opposed to this shallow, n-gram-based approach, here, the internal structure of utterances is exploited: based on semantic considerations, syntactic patterns have been devised that extract triples of event pairs and relation words. The resulting BKB provides a distributional approximation of the discourse relations that can hold between two specific event types. Both approaches exploit complementary sources of knowledge, and may be combined with each other to achieve a more precise prediction of implicit discourse connectives.

The validity of the approach was evaluated with respect to three evaluation criteria: The extracted associations between relation words and event pairs could be shown to be statistically significant, and to be reproducible on other corpora; for three highly frequent relation words, theoretical predictions about their relative distribution could be confirmed, indicating their interpretability in terms of presupposed taxonomies of discourse relations.

Another prospective field of application can be seen in NLP applications, where selection preferences for relation words may serve as a cheap replacement for full-fledged discourse parsing. In the Natural Language Understanding domain, the BKB may help to disambiguate or to identify discourse relations between different events; in the context of Machine Translation, it may represent a factor guiding the insertion of relation words, a task that has been found to be problematic for languages that differ in their inventory and usage of discourse markers, e.g., German and English (Stede and Schmitz 2000). The approach is language-independent (except for the syntactic extraction patterns), and it does not require manually annotated data. It would thus be easy to create background knowledge bases with relation words for other languages or specific domains – given a sufficient amount of textual data.

Related research includes, for example, the unsupervised recognition of causal and temporal relationships, as required, for example, for the recognition of textual entailment. Riaz and Girju (2010) exploit distributional information about pairs of utterances. Unlike approach described here, they are not restricted to adjacent utterances, and do not rely on explicit and recurrent relation words. Their approach can thus be applied to comparably small data sets. However, they are restricted to a specific type of relations whereas here the entire bandwidth of discourse relations that are explicitly realized in a language are covered. Prospectively, both approaches could be combined to compensate their respective weaknesses.

Similar observations can be made with respect to Chambers and Jurafsky (2009) and Kasch and Oates (2010), who also study a single discourse relation (narration), and are thus more limited in scope than the approach described here. However, as their approach extends beyond pairs of events to complex event chains, it seems that both approaches provide complementary types of information and their results could also be combined in a fruitful way to achieve a more detailed assessment of discourse relations.

The goal of this paper was to evaluate the methdological validity of the approach. It thus represents the basis for further experiments, e.g., with respect to the enrichment the BKB with information provided by Riaz and Girju (2010), Chambers and Jurafsky (2009) and Kasch and Oates (2010). Other directions of subsequent research may include address more elaborate models of events, and the investigation of the relationship between relation words and taxonomies of discourse relations.

Acknowledgments

This work was supported by a fellowship within the Postdoc program of the German Academic Exchange Service (DAAD). Initial experiments were conducted at the Collaborative Research Center (SFB) 632 "Information Structure" at the University of Potsdam, Germany. I would also like to thank three anonymous reviewers for valuable comments and feedback, as well as Manfred Stede and Ed Hovy whose work on discourse relations on the one hand and proposition stores on the other hand have been the main inspiration for this paper.

References

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

N. Chambers and D. Jurafsky. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 602–610. Association for Computational Linguistics, 2009.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54, 2008.

Morton Ann Gernsbacher, Rachel R. W. Robertson, Paola Palladino, and Necia K. Werner. Managing mental representations during narrative comprehension. *Discourse Processes*, 37(2):145–164, 2004.

N. Kasch and T. Oates. Mining script-like structures from the web. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 34–42. Association for Computational Linguistics, 2010.

A. Knott and R. Dale. Using linguistic phenomena to motivate a set of coherence relations. *Discourse processes*, 18(1):35–62, 1994.

J. van Kuppevelt and R. Smith, editors. *Current Directions in Discourse and Dialogue*. Kluwer, Dordrecht, 2003.

William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

J. Nivre, J. Hall, and J. Nilsson. Maltparser: A data-driven parser-generator for dependency parsing. In *Proc. of LREC*, pages 2216–2219. Citeseer, 2006.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The penn discourse treebank 2.0. In *Proc. 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, 2008.

M. Riaz and R. Girju. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 361–368. IEEE, 2010.

M. Stede and B. Schmitz. Discourse particles and discourse functions. *Machine translation*, 15(1): 125–147, 2000.

Enric Vallduví. *The Informational Component*. Garland, New York, 1992.

Bonnie L. Webber. Structure and ostension in the interpretation of discourse deixis. *Natural Language and Cognitive Processes*, 2(6):107–135, 1991.

Bonnie L. Webber, Matthew Stone, Aravind K. Joshi, and Alistair Knott. Anaphora and discourse structure. *Computational Linguistics*, 4(29):545–587, 2003.

Z.-M. Zhou, Y. Xu, Z.-Y. Niu, M. Lan, J. Su, and C.L. Tan. Predicting discourse connectives for implicit discourse relation recognition. In *COLING 2010*, pages 1507–1514, Beijing, China, August 2010.