Role Semantics for Better Models of Implicit Discourse Relations

Michael Roth
Department of Language Science and Technology, Saarland University
mroth@coli.uni-sb.de

Abstract

Predicting the structure of a discourse is challenging because relations between discourse segments are often implicit and thus hard to distinguish computationally. I extend previous work to classify implicit discourse relations by introducing a novel set of features on the level of semantic roles. My results demonstrate that such features are helpful, yielding results competitive with other feature-rich approaches on the PDTB. My main contribution is an analysis of improvements that can be traced back to role-based features, providing insights into why and when role semantics is helpful.

1 Introduction

Understanding natural language texts involves, inter alia, correctly identifying coherent segments and the relations that hold between them. Recognizing discourse relations is an important part of this process because such relations not only conceptualize which parts of a text belong together but also *how* they are related. Apart from direct applications in text analysis (e.g., discourse parsing), recognizing discourse relations has further proven a useful preprocessing step for a range of downstream tasks (Louis et al., 2010; Guzmán et al., 2014; Narasimhan and Barzilay, 2015; Chandrasekaran et al., 2017, inter alia).

From a computational perspective, it has been shown that recognizing discourse relations can be performed with high accuracy when explicit discourse markers are available (Pitler et al., 2008). However, classifying relations without explicit markers, so-called *implicit* discourse relations, has persisted as a difficult task to date (cf. Xue et al., 2016). One of the main challenges, as identified in Lin et al. (2009), is the need to perform inference over two discourse segments. In this paper, I propose a new set of features based on semantic roles to address this challenge. These features are meant to provide a shallow form of semantic representation, which might help a classifier to make better informed classification decisions. I argue that role semantic representations are particularly well-suited for this task because different types of discourse relations are defined over the propositions that they connect. For example, definitions in the Penn Discourse TreeBank 2.0 annotation manual (Prasad et al., 2007) explicitly refer to role-level concepts such as events, situations and involved participants. In Rhetorical Structure Theory (Mann and Thompson, 1988), some definitions contain references to concepts akin to *proto-roles* (e.g. "someone's deliberate action"). To illustrate the usefulness of role semantics for the classification of implicit discourse relations, consider the two sentenes shown in Example (1):

(1) a. "Mr. Brady phoned Mr. Greenspan, . . . "
b. "He continued to work the phones through the weekend."
Relation: then, *Temporal.Asynchronous.Precedence* (source: `wsj_2413.pdtb`)

In terms of frame-semantic representation (Fillmore, 1976), the roles involved in the second sentence can be identified as an `Ongoing_activity` (the argument of "continue"), a definite `Duration` and a pronominal `Agent`.[1] These cues indicate a sequence of situations with the same actor, making it likely that a *Temporal* relation holds to the previous sentence.

[1] Roles based on FrameNet, see `http://framenet.icsi.berkeley.edu/`.

2 Discourse Relation Classification with Feature-rich Models

The task addressed in this paper is to determine the discourse relations that hold between two implicitly related discourse segments. In this section, I introduce a combined model for this task that aggregates outputs from multiple simpler models (2.1), each of which uses only one type of feature. I then introduce new feature sets based on semantic roles (2.2).

2.1 Model and Previous Features

My motivation for a model combination derives from the observation that different types of features from the literature greatly vary with respect to the associated number of feature instances and how well they generalize. Consequently, there is no unique set of hyperparameters (e.g. level of regularization, thresholding) that works best for all feature types. The proposed combined model consists of two steps to make use of information from inherently diverse feature types. First, I train simple discourse relation classifiers that only use one feature type each. Outputs from multiple classifiers are then combined using averaging as a simple but effective form of model combination.[2]

I formalize the classification of an instance i with respect to a discourse relation r as follows. Given a set of n feature types, feature values are extracted and a set of simple classifiers $c_{r,1} \ldots c_{r,n}$ are trained. At test time, each classifier outputs an individual score $score_{c_{r,j}}(i) \in [0, 1]$. Decisions of multiple classifiers are then aggregated by computing the arithmetic mean of the individual scores. As single classifiers, I use logistic regression models with L2 loss, as implemented in the LIBLINEAR toolkit (Mu-Chu et al., 2015). Accordingly, the aggregated model predicts a relation r for instance i iff $\frac{1}{n}\Sigma_{j=1\ldots n} score_{c_{r,j}}(i) > 0.5$.

The following list provides an overview of all feature sets from the literature that I reimplemented for the described approach, and gives the total number of features for each type.

First/Last. Set of indicators for the first and last words in each discourse segment. In case of Example (1), instances of this feature set include `1:FIRST:Mr.`, `2:FIRST:He`, etc. (for details, see Pitler et al., 2009). ca. 74 000 features

Dates and number. Indicator features for the number of date and number expressions in each discourse segment (e.g. `1:DATE:0`; see Pitler et al., 2009). ca. 10 000

Production rules. Features on production rules used to construct each discourse segment's constituency tree (e.g. `1:S_NP_VP`; see Lin et al., 2009). ca. 78 000

Verb features. Indicators for the main verb, its tense/modality and average verb phrase length (e.g. `1:VERB:phone`, `2:TENSE:past`; see Park and Cardie, 2012). ca. 20 000

Coreference. Set of features that indicate coreferring mentions, as predicted by Stanford CoreNLP (Lee et al., 2013), across two related discourse segments (see Rutherford and Xue, 2014). ca. 10 000

Brown clusters. Feature sets indicating precomputed Brown cluster IDs (Turian et al., 2010) of words occurring in each discourse segment (e.g. `2:11100110`; see Braud and Denis, 2015). 200–6 400

Pairwise Brown clusters. Pairwise Brown cluster IDs indicating word pairs across two related discourse segments (e.g. `11110110x11000100`; see Braud and Denis, 2015). up to 10 million

[2] Sum/averaging is used here because of its simplicity and robustness (Kittler et al., 1998). Due to the small development set size, methods with additional parameters may tend to overfit.

2.2 Features based on Semantic Roles

As new features, I propose to utilize the semantic roles identified in a pair of discourse segments. I define two variants of this feature type: one based on FrameNet (Ruppenhofer et al., 2010) and one based on PropBank (Palmer et al., 2005). All features are computed automatically using a state-of-the-art semantic role labeler (Roth, 2016; Roth and Lapata, 2016). Each variant includes both raw labels as well as a combination of the label and the filler word to which the label is assigned. To reduce sparsity, filler words are always represented by pre-computed Brown cluster IDs (Turian et al., 2010). The list below provides additional details as well as example instances based on the sentences shown in Example (1).

FrameNet roles. This feature set indicates all frame elements that are identified in a pair of related discourse segments. For instance, two frame element fillers are identified in the phrase *he continued to work*: *he* is the `Agent` of the frame evoked by the verb *work*, and *work* itself fills the `Ongoing_activity` element of the frame evoked by *continue*. To compute features for *he*, the Brown cluster ID of the word is looked up (`11100110`) and it is determined that the word occurs in the `2`nd discourse segment in Example (1). Accordingly, the indicator features that represent *he* and its semantic role in this case are `2:Agent` and `2:Agent:11100110`.[3] ca. 37 000

PropBank roles. Analogous to the FrameNet features, this feature set consists of indicators for PropBank labels. Because argument labels in PropBank (A0. . . A5) are only meaningful with respect to a given predicate, I define two conjoined versions of this feature type: one takes into account the predicate's class in VerbNet (Kipper et al., 2008) and one the predicate lemma itself (e.g., `2:work-73.2_A0` and `2:work_A0:11100110`, resp.). In each variant, predicate-independent labels (modifiers such as time and location) are optionally considered in the same representation format. ca. 560 000

3 Experiments

I evaluate the proposed model on version 2.0 of the Penn Discourse Treebank (PDTB, Prasad et al., 2008). To ensure a fair comparison, I use the same preprocessing and weighting techniques as well as the same data instances as previous work (Rutherford and Xue, 2014; Braud and Denis, 2015). That is, each instance is a pair of implicitly related discourse segments as annotated in the PDTB corpus. Sections 2–20 of the corpus are used for training, 21–22 for testing, and all other sections for development.

Baseline and comparison models. I use three variants of the proposed model to directly examine the utility of semantic roles and combining classifiers. The first two models are instances of the feature-rich model described in Section 2, with hyperparameter tuning and feature selection done on the training and development sets: *AverageFeats* uses a combination of feature sets described in subsection 2.1, whereas *AverageFeats+SRL* also uses the role-level features from subsection 2.2. Note that for each type of role set at most one feature representation is chosen. All feature sets are selected based on the best performance on the development set. The third model, *AllFeats*, is a baseline logistic regression classifier that uses all best development feature sets at the same time.

For comparison, I consider a range of current state-of-the-art models. The best feature-rich models (Rutherford and Xue, 2014; Braud and Denis, 2015) use a range of binary indicator features largely identical to the features described in Section 2.1. The most notable difference to this work is that Rutherford and Xue use a small list of coreference patterns in addition to features that simply indicate coreferring mention counts. Neural-network models (Zhang et al., 2015; Liu and Li, 2016; Qin et al., 2016) use attention or convolution mechanisms to identify important words and word spans in each discourse segment. They then predict the discourse relation based on a composition function applied over representations of important words. All of the comparison models use the same training and test instances as this work and are directly comparable.

[3] I also experimented with feature conjunctions in order to explicitly model semantic interactions between two discourse segments. However, such conjuctions consistently reduced development performance, probably due to sparsity.

	comp	cont	exp	temp
Neural network models				
Zhang et al. (2015)	33.2	52.0	69.6	30.5
Liu and Li (2016)	36.7	54.5	70.4	**38.8**
Qin et al. (2016)	**41.6**	**57.3**	**71.5**	35.4
Recent feature-rich models				
Rutherford and Xue (2014)	39.7	54.4	70.2	28.7
Braud and Denis (2015)	36.4	55.8	67.4	29.3
This work's models				
AverageFeats	36.3	55.9	69.4	30.5
AverageFeats+SRL	37.0	56.3	69.4	32.1
AllFeats	34.5	51.3	60.4	26.8

Table 1: One-vs-all results in F_1-score on the four PDTB top-level relations (*comp*arison, *cont*ingency, *exp*ansion and *temp*oral). Best overall results are marked in bold, best results by feature-rich models are underlined.

Role name	Position	Weight
Request	segment 1	+1.13061
Addressee	segment 1	+0.90852
Relative_time	segment 1	+0.85555
Stuff	segment 2	+0.79267
Success_or_failure	segment 2	+0.69008
Unattr_information	segment 2	+0.66578
Agent	segment 1	+0.39992
Agent	segment 2	−0.68378

Table 2: List of indicator features on FrameNet frame elements that received a high weight for recognizing the discourse relation Contingency.

Results. Table 1 lists F_1-scores for each of the top-level relations in the PDTB test set. Note that multiple relation types can apply to one relation instance. Hence, instead of one 4-way classification, this task is traditionally separated into four binary tasks. The results show that *AverageFeats* performs competitively with other feature-rich models for discourse relation classification. Additional features on semantic roles improve performance for all but one relation. In the cases in which semantic roles are helpful, both FrameNet-based and PropBank-based feature sets are selected. Two of the four scores by *AverageFeats+SRL* represent the best reported results with a feature-rich model. The performance of *AllFeats* is consistently worse than those of other recent models. This complies with my hypothesis that hyperparameters tuned for one single model do not generalize well across different feature types.

Discussion. One advantage of simple classification models based on binary features is that predictions based on learned feature weights can easily be interpreted. In the following, I take a closer look at classification instances that the model *AverageFeats+SRL* got correct but that were misclassified by the other models. The weights of the features that apply in these examples provide insights as to how and when semantic roles are beneficial. For simplicity, I focus the discussion on FrameNet roles (i.e. frame element types).

For the implicit relation *Contingency*, the learned feature weights indicate that its prediction becomes more likely when an `Agent` is identified in the first discourse segment (high positive feature weight) but not in the second segment (negative feature weight). This seems to reflect the fact that most of these relations connect a cause and a result, as shown for instance in Example (2).

(2) "... traders can buy or sell even when they don't have a customer order ... [*as a result*] liquidity becomes a severe problem for thinly traded contracts ..." (`wsj_2110.pdtb`)

Semantic roles are helpful in such cases because they provide a means to distinguish events initiated by someone (the cause) from simple states (the result). A list of features that seem to contribute to this distinction, as identified by their associated feature weights, are given in Table 2.

The feature weights assigned in role-based classifiers for other discourse relations are overall smaller and thus harder to interpret. Still, certain trends can be observed. For example, I find that co-occurrences of specific roles in both connected discourse segments may indicate a *Comparison*. Example (3) shows one such instance, in which the role `Purpose` has been identified in both segments (assigned feature

weight: $+0.117$). Other roles, for which the same pattern of weights are observed include, among others, `Theme` ($+0.435$) and `Businesses` ($+0.254$).

(3) "Her goal: to top 300 ad pages ... [*but*] whether she can meet that ambitious goal is still far from certain." (`wsj_2109.pdtb`)

Concerning the *Temporal* relation, high feature weights are learned for specific FrameNet roles, such as `Activity_start` in the first discourse segment ($+1.654$) and `Process_end` in the second segment ($+1.116$). Even though these feature weights seem to be intuitive, they only lead to marginal improvements to the absolute classification performance, presumably because textual order in discourse not necessarily represents linear temporal order ("before" vs. "after"). Higher gains could be achieved if training and evaluation was performed on more specific relation annotations but such instances are too rare in practice for the feature-rich classifiers to learn robust generalizations: For example, the current version of the Penn Discourse Treebank contains a total of only 151 implicit relation instances of the discourse relation *Temporal.Asynchronous.Succession*.

4 Related Work

The task of predicting implicit discourse relations was first introduced in the context of implicit and explicit relation classification (Marcu and Echihabi, 2002). Pitler et al. (2009) were the first to address implicit relations specifically. They applied a Naive Bayes model with a range of binary features. Follow-up work examined different methods for feature selection (Lin et al., 2009; Park and Cardie, 2012) as well as novel feature types based on pairs of word classes/clusters, entity mentions, and word embeddings (Biran and McKeown, 2013; Louis et al., 2010; Braud and Denis, 2015). Further improvements were made via multi-task learning (Lan et al., 2013) and training data expansion (Rutherford and Xue, 2015).

In recent years, a myriad of neural-network based models have been proposed for the task of recognizing implicit discourse relations (Ji and Eisenstein, 2014; Zhang et al., 2015; Liu and Li, 2016; Qin et al., 2017, inter alia). Models of this kind have a high expressive power and generally outperform methods that rely on manual feature engineering. However, being able to trace back improvements to individual features was key to my discussion in Section 3. Recent results in downstream NLP tasks indicate that neural network models can perform better when incorporating binary features (Cheng et al., 2016; Sennrich and Haddow, 2016, inter alia).

5 Conclusions

I proposed a simple model combination for discourse relation classification that aggregates outputs from multiple classifiers. Several classifiers use novel features based on automatic semantic role labeling. I have shown that such features improve classification performance and provide shallow insights into relationships between role semantics and discourse semantics.

In the future, I plan to apply more sophisticated methods of model ensembling. I would like to investigate whether neural network approaches to discourse relation classification can also benefit from structural information in the form of semantic roles. I believe this to be a promising research direction especially because of the small size of available training data, which presumably makes it difficult for a neural network to learn any higher level structures by itself.

Acknowledgements

This research was supported in part by the Cluster of Excellence "Multimodal Computing and Interaction" of the German Excellence Initiative, and a DFG Research Fellowship (RO 4848/1-1).

References

Biran, O. and K. McKeown (2013). Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, pp. 69–73.

Braud, C. and P. Denis (2015). Comparing word representations for implicit discourse relation classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 2201–2211.

Chandrasekaran, M. K., C. Demmans Epp, M.-Y. Kan, and D. Litman (2017). Using discourse signals for robust instructor intervention prediction. In *31st AAAI Conference on Artificial Intelligence*, San Francisco, California. to appear.

Cheng, H.-T., L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al. (2016). Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, Boston, Massachusetts, pp. 7–10.

Fillmore, C. J. (1976). Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, Volume 280, pp. 20–32.

Guzmán, F., S. Joty, L. Màrquez, and P. Nakov (2014). Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, pp. 687–698.

Ji, Y. and J. Eisenstein (2014). Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, pp. 13–24.

Kipper, K., A. Korhonen, N. Ryant, and M. Palmer (2008). A large-scale classification of english verbs. *Language Resources and Evaluation Journal 42*(1), 21–40.

Kittler, J., M. Hatef, R. P. Duin, and J. Matas (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*(3), 226–239.

Lan, M., Y. Xu, and Z. Niu (2013). Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria, pp. 476–485.

Lee, H., A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, and D. Jurafsky (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics 39*(4), 885–916.

Lin, Z., M.-Y. Kan, and H. T. Ng (2009). Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 343–351.

Liu, Y. and S. Li (2016). Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, pp. 1224–1233.

Louis, A., A. Joshi, and A. Nenkova (2010). Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, Tokyo, Japan, pp. 147–156.

Louis, A., A. Joshi, R. Prasad, and A. Nenkova (2010). Using entity features to classify implicit discourse relations. In *Proceedings of the SIGDIAL 2010 Conference*, Tokyo, Japan, pp. 59–62.

Mann, W. C. and S. A. Thompson (1988). Rhetorical structure theory. Toward a functional theory of text organization. *Text 8*(3), 243–281.

Marcu, D. and A. Echihabi (2002). An unsupervised approach to recognizing discourse relations. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, pp. 368–375.

Mu-Chu, L., C. Wei-Lin, and L. Chih-Jen (2015). Fast matrix-vector multiplications for large-scale logistic regression on shared-memory systems. In *IEEE International Conference on Data Mining*, Atlantic City, New Jersey.

Narasimhan, K. and R. Barzilay (2015). Machine comprehension with discourse relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China, pp. 1253–1262.

Palmer, M., D. Gildea, and P. Kingsbury (2005). The Proposition bank: An annotated corpus of semantic roles. *Computational Linguistics 31*(1), 71–106.

Park, J. and C. Cardie (2012). Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Seoul, South Korea, pp. 108–112.

Pitler, E., A. Louis, and A. Nenkova (2009). Automatic sense prediction for implicit discourse relations in text. 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*, Suntec, Singapore, pp. 683–691.

Pitler, E., M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, and A. Joshi (2008). Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, Manchester, United Kingdom, pp. 87–90.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-2008)*, Marrakesh, Marocco.

Prasad, R., E. Miltsakaki, N. Dinesh, A. Lee, A. Joshi, L. Robaldo, and B. Webber (2007). The penn discourse treebank 2.0 annotation manual. Technical report.

Qin, L., Z. Zhang, and H. Zhao (2016). A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, pp. 2263–2270.

Qin, L., Z. Zhang, H. Zhao, Z. Hu, and E. Xing (2017). Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, pp. 1006–1017. Association for Computational Linguistics.

Roth, M. (2016). Improving frame semantic parsing via dependency path embeddings. In *Book of Abstracts of the 9th International Conference on Construction Grammar*, Juiz de Fora, Brazil, pp. 165–167.

Roth, M. and M. Lapata (2016). Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, pp. 1192–1202.

Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck, C. R. Johnson, and J. Scheffczyk (2010). FrameNet II: Extended Theory and Practice. Technical report, International Computer Science Institute.

Rutherford, A. and N. Xue (2014). Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, pp. 645–654.

Rutherford, A. and N. Xue (2015). Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, pp. 799–808.

Sennrich, R. and B. Haddow (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany, pp. 83–91.

Turian, J., L.-A. Ratinov, and Y. Bengio (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 384–394.

Xue, N., H. T. Ng, S. Pradhan, A. Rutherford, B. Webber, C. Wang, and H. Wang (2016). Conll 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, Berlin, Germany, pp. 1–19.

Zhang, B., J. Su, D. Xiong, Y. Lu, H. Duan, and J. Yao (2015). Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 2230–2235.