Incorporating annotator uncertainty into representations of discourse relations

Anonymous ACL submission

Abstract

Annotation of discourse relations is a known difficult task, especially for non-expert annotators. In this paper, we investigate novice annotators' uncertainty on the annotation of discourse relations on spoken conversational data. We find that dialogue context (single turn, pair of turns within speaker, and pair of turns across speakers) is a significant predictor of confidence scores. We compute distributed representations of discourse relations from co-occurrence statistics, and incorporate information about confidence scores and dialogue context. We perform a hierarchical clustering analysis using these representations and show that weighting discourse relations representations with information about confidence and context kind coherently models our annotators' uncertainty about discourse relations labels.

1 Introduction

Discourse relations (DRs) are those relations such as Elaboration, Explanation, Narration, which hold between discourse units and give coherence to texts. The task of labeling DRs is known to pose difficulties for annotators (Spooren and Degand, 2010), as sometimes more than one interpretation is possible, or more than one discourse relation may hold (Scholman et al., 2022; Webber, 2013).

Recent studies have shown that allowing for multiple labels in annotation can improve the performance of discourse parsers (Yung et al., 2022). Scholman et al. (2022) test different label aggregation methods and find that probability distribution labels better capture ambiguous interpretations of discourse relations. (1) shows an example from their corpus, where the relation between the second and third sentences is annotated with both Conjunction and Result.

(1) It is logical that our attention is focused on cities. Cities are home to 80% of the 500 million or so inhabitants of the EU. It is in cities that the great majority of jobs, companies and centres of education are located. (DiscoGeM, Europarl genre; Scholman et al., 2022)

This double-label captures that these two segments provide two conjoined facts about cities, but can also be perceived as holding a causal relation (because cities are home to the largest part of the population, most jobs, companies and educational institutions are located there).

In this work, we investigate which relations are usually perceived as similar or co-occurring in spontaneous conversations by individual annotators. We are particularly interested in how novice annotators interpret discourse relations categories when annotating spoken conversational data. We collect annotations of DRs of Switchboard (Godfrey et al., 1992) telephone conversations together with confidence scores, among novice annotators, allowing for multiple labels. We find that confidence scores vary significantly across dialogue context (single turn vs pair of turn within speaker vs pair of turn across speakers). We incorporate information about dialogue context kinds and confidence scores into distributed representations of discourse relations. A clustering analysis shows that discourse relations that tend to occur across speakers cluster together, as opposed to discourse relations which tend to occur within a speaker, either within or across turn.

2 Annotation of Discourse Relations

We selected 19 conversations[1] from the Switchboard corpus (Godfrey et al., 1992), which contains telephone conversations between pairs of participants about a variety of topics (e.g. recycling, movies, child care). We chose this corpus because it contains informal, spontaneous dialogues, and

[1]We discarded the annotations from one conversation because the annotators did not follow the guidelines.

because it has been used within linguistics in various studies on conversation (Jaeger and Snider, 2013; Reitter and Moore, 2014).

An initial set of turns for annotation was selected by using a dependency parser to select elementary discourse units (EDUs; Asher and Lascarides, 2003) with two or more roots or verbs. Our annotation scheme is a modified version of STAC (Asher et al., 2016) which has been adjusted to accommodate the frequent interruptions, short, and non-clausal nature of many utterances.

Since many EDUs are very short in nature, we selected pairs of individual EDUs and complex EDUs (CDUs) for discourse relation annotation. CDUs consist of two or more EDUs which constitute an argument to a DR (Asher and Lascarides, 2003). We selected items for annotation across three different contexts: within a single turn (within speaker), and across two turns within speaker, and across two immediately adjacent turns (two speakers). (2) shows an example for each of the discourse context kinds, respectively with the first EDU (or CDU) in italics and the second in bold.

(2) a. A: ‖ *and they discontinued them* ‖ **because people were coming and dumping their trash in them.** ‖

b. A: ‖ *We live in the Saginaw area.* ‖
B: ‖ **Saginaw?** ‖

c. B: ‖ No, ‖ *I just, I noticed* ‖ *in Iowa and other cities like that, it's a nickel per aluminum can.* ‖
A: ‖ Oh. ‖
B: ‖ **So you don't see too many thrown out around the ‖ [laughter] ‖ streets**.

2.1 Discourse Relations Taxonomy

The DRs chosen to annotate our corpus were adapted from the STAC corpus manual (Asher et al., 2012). Table 1 shows the taxonomy used. We focused on a small taxonomy because of the novice nature of annotators. Future work will include revising the taxonomy used. We selected 11 DRs based on a pilot annotation of one conversation by one author, and added an "Other" category for relations not included in the list of labels.

2.2 Annotation Procedure

The annotation of discourse relations was done by students enrolled in a Computational Linguistics class. Each conversation was annotated by about 5 students for course credit. Annotators were trained using written guidelines, a quiz-like game, and a live, group annotation demo. We used the annotation interface Prodigy (Montani and Honnibal, 2018). Each display presented the two target discourse units plus two context turns before and two after. Below the text, the screen showed a multiple choice list of discourse relations plus the "Other" category. Lastly, each display asked for confidence scores in the range 1-5, corresponding to least confident to most confident.

Acknowledgement	Elaboration
Background	Explanation
Clarification Question	Narration
Comment	Question-Answer Pair
Continuation	Result
Contrast	Other

Table 1: Discourse relations taxonomy.

3 Dialogue Context as a Predictor of Confidence Scores

First we sought to understand how discourse relations and discourse context influence annotator confidence. The hierarchical nature of our data and the ordinal nature of confidence scores make it most appropriate to analyze the data with a Cumulative Link Mixed Model (CLMM; Liddell and Kruschke, 2018) to simultaneously account for individual variability in confidence with respect to participants and EDU pairs in the annotation task. We first built a null model containing only random intercepts by annotator and compared it to a model containing an additional fixed effect and random slope by annotator for discourse context type (*kind*, dummy coded). A likelihood ratio test revealed a significant improvement in fit by adding kind as a predictor ($\chi^2(7) = 126.64, p < 0.001$). Adding random intercepts for EDU pairs to account for annotation difficulty across EDU pairs also led to a significant improvement in model fit beyond the model containing discourse context kind ($\chi^2(1) = 195.01, p < 0.001$). This suggests that our annotators' confidence scores are sensitive to the context of EDU pairs.

Figure 1 shows mean confidence scores per context kind across discourse relations. Confidence scores within a speaker both across and within turns received similar confidence ratings ($\beta = -0.13$,

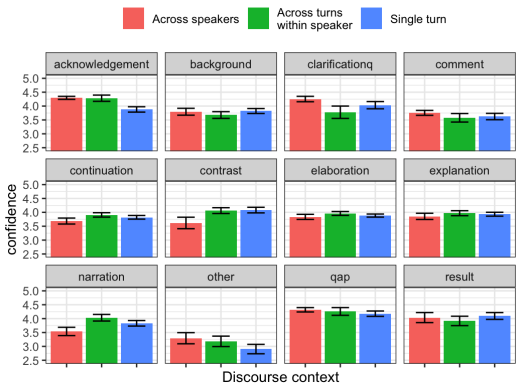


Figure 1: Confidence scores per context kind across discourse relations. *qap* stands for Question-Answer Pair and *clarificationq* for Clarification Question.

$z = -0.56$, $p =$ n.s.), while annotators were significantly more confident for relation annotation across speakers ($\beta = 0.63$, $z = 3.05$, $p < .01$). The CLMM revealed that annotators used confidence scores between 3 and 5 overall, except for the label "Other", for which they selected lower confidence scores. Background received lower confidence scores overall. Continuation, Contrast and Narration received higher scores for contexts within speaker. Comment and Result received higher scores for turns across speakers and single turn. For Elaboration and Explanation, mean confidence scores are very similar across the three contexts, with slightly higher scores for single turn and pairs of turns within speaker. Acknowledgment, Clarification Question ("clarificationq") and Question-Answer Pair ("qap") received higher scores for turns across speakers, which makes sense given the dialogic nature of these relations. However, these relations also received rather high confidence scores for single turn and pairs of turns within speaker, which is a bit surprising. We suspect this might be due to including more context than just the relevant EDUs, which might have led annotators to choose relations that did not strictly apply to the pair of highlighted EDUs. Future analysis will look closer at this aspect.

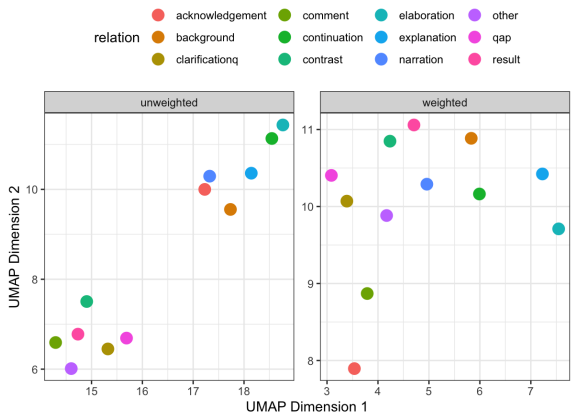
4 Distributed representations from DRs annotations

To model discourse relations similarity as perceived by annotators, we computed embedding representations of discourse relations. We extracted each n individual annotation containing **r**elation-**c**onfidence (r, c) tuples selected by a given annotator for a pair of EDUs. We combine bag-of-relations vectors with one-hot encoded features representing the discourse context kind and multiply the counts of relations (within annotator, either 1 or 0 for each relation) by the confidence score (1-5). This weighting emphasizes high confidence; an ideal reweighting may be possible with additional parameter search, possibly in conjunction with the CLMM outputs.

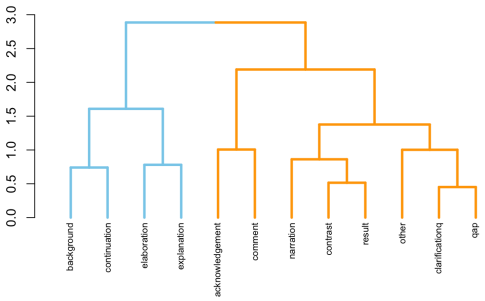
For an $n \times 1$ confidence ratings matrix C, an $n \times 12$ bag-of-relations matrix R, an $n \times 3$ discourse context matrix D for each annotation, we obtain an Annotation Matrix $A = C \times (R|D)$. We then obtain a co-occurrence matrix O such that $O = A \cdot A^T$, which we factorize without an intercept shift using Principal Component Analysis following Levy and Goldberg (2014). Each relation is thus represented as a vector that consolidates co-occurrences between all relations within a single annotator that are weighted by confidence score.

We then projected these embeddings into two dimensions with UMAP (McInnes et al., 2018) and performed a hierarchical clustering analysis over the resulting coordinates, which we visualized in Figure 2a. Figure (2b) shows a dendrogram with the output clusters, colored according to the optimal number of clusters ($k = 2$), calculated using average silhouette widths (Levshina, 2022). There are two large clusters, one of which contains two sub-clusters with Background and Continuation, on the one hand, and Elaboration and Explanation on the other. In the other large cluster, Acknowledgement and Comment form a sub-cluster. These are very common relations between pairs of turns across speakers. Clarification Question and Question-Answer Pair form another sub-cluster, also common relations between pairs of turns across speakers, in close proximity to the Other label, which received a sub-cluster of its own. Narration and Contrast and Result, form the last sub-clusters, which we suspect is due in part to the frequencies of these relations (Schnabel et al., 2015). We include a dendrogram with the output clusters of a hierarchical clustering analysis performed with base bag-of-relations vectors (without context kind and confidence scores weight) in Figure 3 in the Appendix for comparison.

Currently, we provide these results as a proof of concept of the feasibility and interpretability of noisy labels produced by novice annotators. Importantly, annotations weighted by confidence produce coherent clusters of discourse relations. We



(a) The coordinates obtained with UMAP for all discourse relations plotted in two-dimensional space. The plot on the left shows the unweighted embedding representations and the figure on the right shows the weighted embedding representations.



(b) Dendrogram showing hierarchical clustering of Discourse Relations built from UMAP coordinates. *qap* stands for Question-Answer Pair and *clarificationq* for Clarification Question.

Figure 2: Dimensionality reduction and clustering of relation embeddings.

envision applications of DR embeddings to several domains including dialogue generation, such that appropriate responses to input are partially conditioned on a latent or mixed combination of DRs.

5 Related Work

Discourse relations annotation is usually done within Rhetorical Structure Theory (Mann and Thompson, 1987), as in the RST-DT (Carlson et al., 2003) and GUM (Zeldes et al., 2021) corpora, within Segmented Discourse Representation Theory (SDRT, Asher and Lascarides, 2003), as in the STAC (Asher et al., 2016) and Molweni (Li et al., 2020) corpora, or within the Penn Discourse Treebank framework (Prasad et al., 2008, 2018). We use a taxonomy adapted from SDRT, in particular, the STAC corpus, because we found that it adapts well to conversational data.

Annotators are usually trained to identify discourse relations using the framework's taxonomy. Some recent alternatives to explicitly collecting annotation of DRs include crowdsourcing by eliciting connectives (Yung et al., 2019; Scholman et al., 2022) or question-answer pairs (Pyatkin et al., 2020) rather than relations. In this work, we wanted to investigate how annotators perceive discourse relations categories, and therefore a connective insertion task would only provide indirect evidence. We train annotators on DRs labeling and ask annotators to choose from a set of discourse relations labels. We allow for multiple labels to investigate what relations are more confusable or perceived as co-occurring (Marchal et al., 2022).

6 Discussion and Future Work

In this study, we collected multiple discourse relations annotations from a subset of the Switchboard corpus, together with confidence scores. We found that dialogue context had a significant effect on confidence scores. We computed embedding representations of DRs using co-occurrence statistics and weighted the vectors using context type and confidence scores, and found that these representations coherently model our annotators uncertainty about discourse relations labels.

In the future, we plan to use this information to improve our annotation guidelines and training process in order to run a larger scale annotation study of the Switchboard corpus to analyze discourse relations patterns in spoken dialogues. Additionally, we are interested in comparing these latent representations of discourse relations with a categorization study in which no labels are presupposed (e.g. an EDU pairs pile sorting task), to investigate whether these representations resemble how language users perceive and categorize relations between discourse segments.

Limitations

This work is limited by the size of the dataset and the taxonomy used in the annotation task. While we found that our annotators perceived some of the categories as more similar or confusable, future work can examine annotators' uncertainty in a

larger set of discourse relations.

Additionally, we focused on uncertainty in novice annotators. Annotation and selection of confidence scores by expert annotators might yield different results.

Ethics Statement

We are not aware of ethical issues associated with the texts used in this work. Students participated in the annotation task as part of course credit but annotation decisions were not associated with their performance in the course.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Nicholas Asher, Vladimir Popescu, Philippe Muller, Stergos Afantenos, Anais Cadilhac, Farah Benamara, Laure Vieu, and Pascal Denis. 2012. Manual for the analysis of settlers data. *Strategic Conversation (STAC). Université Paul Sabatier*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

T Florian Jaeger and Neal E Snider. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, 127(1):57–83.

Natalia Levshina. 2022. Semantic maps of causation: New hybrid approaches based on corpora and grammar descriptions. *Zeitschrift für Sprachwissenschaft*, 41(1):179–205.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *NIPS*.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. *arXiv preprint arXiv:2004.05080*.

Torrin M Liddell and John K Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79:328–348.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing annotation quality in multi-label annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Ines Montani and Matthew Honnibal. 2018. Prodigy: A new annotation tool for radically efficient machine teaching. *Artificial Intelligence*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the pdtb: The next generation. In *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. Qadiscourse–discourse relations as qa pairs: Representation, crowdsourcing and baselines. *arXiv preprint arXiv:2010.02815*.

David Reitter and Johanna D Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Merel Scholman, Dong Tianai, Frances Yung, and Vera Demberg. 2022. Discogem: A crowdsourced corpus of genre-mixed implicit discourse relations. In *the 13th Language Resources and Evaluation Conference*

(LREC 2022), pages 3281–3290. European Language Resources Association.

Wilbert Spooren and Liesbeth Degand. 2010. Coding coherence relations: Reliability and validity.

Bonnie Webber. 2013. What excludes an alternative in coherence relations? In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 276–287, Potsdam, Germany. Association for Computational Linguistics.

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

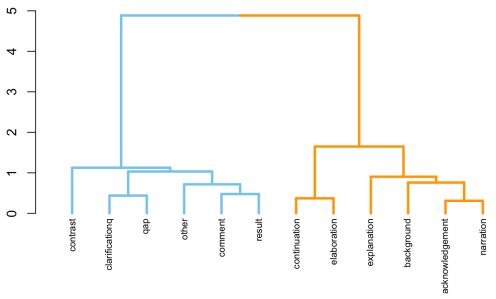


Figure 3: Dendrogram showing hierarchical clustering of Discourse Relations built from UMAP coordinates without context kind and confidence scores weighting. *qap* stands for Question-Answer Pair and *clarificationq* for Clarification Question.