A Weakly-Supervised Learning Approach to the Identification of "Alternative Lexicalizations" in Shallow Discourse Parsing

René Knaebel
Applied Computational Linguistics
Department of Linguistics
University of Potsdam
Germany
rene.knaebel@uni-potsdam.de

Abstract

Recently, the identification of free connective phrases as signals for discourse relations has received new attention with the introduction of statistical models for their automatic extraction. The limited amount of annotations makes it still challenging to develop well-performing models. In our work, we want to overcome this limitation with semi-supervised learning from unlabeled news texts. We implement a self-supervised sequence labeling approach and filter its predictions by a second model trained to disambiguate signal candidates. With our novel model design, we report state-of-the-art results and in addition, achieve an average improvement of about 5% for both exactly and partially matched alternatively–lexicalized discourse signals due to weak supervision.

1 Introduction

Understanding the underlying structure of a text is a fundamental problem in computational linguistics. In discourse analysis, shallow discourse parsing in particular, we aim to identify individual discourse relations within a text. Thus we can gain information that helps in downstream tasks such as automatic summarization, machine translation, and document classification. The study of *connecting phrases* not only helps in understanding the way people connect their thoughts but also in the identification of discourse relations anchored by them. For our work, we use the third version of the **Penn Discourse Treebank** (PDTB) (Prasad et al., 2018) that distinguishes between **explicit** relations (signaled by a closed set of discourse connectives, e.g *because*, *and*, *if-then*, and *before*) and **alternative lexicalizations** (signaled by connecting phrases other than discourse connectives, e.g. *this means*,*for that reason*, and *it all adds up to*). In total, the PDTB contains 25878 signaled relations, most of which belong to the group of explicit relations (94%). Only 1638 connecting phrases build the group of free connective phrases, in the corpus referred to as alternative lexicalizations. While explicit relations are more commonly used to verbalize expansions and comparisons between text spans, alternative lexicalizations often point to lexically grounded causal relations. Also, they potentially contain information, e.g. the phrase *the most crucial reason for that* gives also evidence about the reason's importance, which is useful for understanding the full discourse.

In our work, we aim to overcome the problem of very limited training data available for free connective phrases and examine a weakly-supervised scenario for continuously improving a model through its own predictions. We regularize these predictions by re-ranking the extracted signals through a separate model trained to discriminate possible signal candidates into signals with or without discourse usage. Summarized, our contributions are: We (i) present a novel architecture and provide state-of-the-art results for recognizing alternative lexicalizations in the recent version of the PDTB. Further, we (ii) improve its performance of recognizing phrases by integrating unlabeled data into the training process using weak supervision.

2 Related Work

Self-supervised learning (Yarowsky, 1995), the most simple semi-supervised learning algorithm, extends its training data by adding new samples with confident predictions on different data. Self-training has been successfully applied on constituent parsing (McClosky et al., 2006) by incorporating a re-ranking strategy (Charniak and Johnson, 2005; Collins and Koo, 2005) to improve parsing results and reduce the bias of the trained model. Also, Suzuki and Isozaki (2008) improved performance on part-of-speech tagging via sequence labeling. In recent work, Nishida and Matsumoto (2022) study the empirical effectiveness of bootstrapping annotations from out-of-domain data and show its positive impact for BERT-based discourse dependency parsers. For candidates selection, they study criteria inspired by Steedman et al. (2003).

Chou et al. (2014) approach semi-supervised learning for named entity recognition, a similar training problem (sequence labeling) as ours. They propose an additional model for estimating confidence (self-testing) and removing samples with low scores. Braud et al. (2016) first apply semi-supervised learning to RST discourse parsing using multiple views on the data by incorporating various auxiliary tasks, such as PDTB discourse parsing. Knaebel and Stede (2020b) improved their argument extraction by jointly training three separate models so-called tri-training on additional news documents. Recently, Kobayashi et al. (2021) successfully bootstrapped RST sub-trees using a combination of simpler feature-based teachers to train a more complex neural student.

The group of alternative lexicalized relations has been rarely studied. Prasad et al. (2010) did initial work on the identification and analysis of alternative lexicalized relations in an older PDTB version. Synková et al. (2017); Rysová and Rysová (2015) distinguished two classes of alternative lexicalizations and developed a dictionary approach for more regular alternative lexicalized phrases. Most recently, Knaebel and Stede (2022) implemented the first automatic neural-based model for recognizing alternative lexicalizations on a sentence level using a binary sequence labeling approach. In this work, we build on their approach and adapt this model to the paragraph level, similar to the explicit connective model of Kurfalı (2020).

3 Method

3.1 Recognizing alternative lexicalizations

The recognition of alternative-lexicalized discourse signals (AltLex) in the PDTB corpus is challenging due to the higher complexity of the phrases when compared to explicit signals for example, and the limited number of training samples. While Knaebel and Stede (2022) predict binary labels (*is-part-of* the signal) on the sentence level, we follow Kurfalı (2020) and integrate more context into the model by training the whole model on the paragraph level. Accessing more context seems unavoidable for improving performance as discourse signals naturally link to phrases outside their sentence. We make use of pre-trained large language models and fine-tune the base model combined with an additional token classification layer on top of it.

Shifting from sentences to paragraphs results in potentially having an arbitrary number of signals. For this purpose, we use a three-class encoding similar to Kurfalı (2020): single signals, e.g. *following*, *resulting*, *not*, and *soon*, multi-word signals, e.g. *for this reason* and *in addition to*, and no signal otherwise. We limited our experiments to continuous signals, e.g. we removed phrases like *the more [. . .], the more*, which removes a minor number of samples but allows for decoding the labeled sequence without redundancy. We did not choose a more complex signal encoding, such as BIOS and BIOES, due to the lack of available training data and the resulting class imbalance.[1]

3.2 Learning from unlabeled data

In this work, we study self-training, which is a very basic but effective semi-supervised learning technique that uses a model's self-estimation to integrate confident predictions from unlabeled data. However, this technique has a high bias due to reinforcing its own false predictions. We overcome this problem by, first, improving the base performance of our signal extractor by building an ensemble of three separately trained models. Second, we follow the idea of McClosky et al. (2006) and introduce a separate model for confidence estimation that not only reduces the bias of a singly self-trained model but also simplifies the determination of a confidence score.

To estimate the model's confidence in its predicted alternatively–lexicalized phrases, referred to as candidates, we design an auxiliary task to disambiguate signal candidates produced by the labeling model. We want to learn to discriminate candidate phrases into those related to an AltLex or not. We adapt previous work on explicit sense classification (Knaebel and Stede, 2020a) to alternative lexicalizations and simultaneously predict whether a possible candidate phrase is used as a discourse signal and if so, we learn to predict its sense. Instead of learning only a single sense level, we jointly learn sense versus no-sense prediction on coarse and fine senses as Long and Webber (2022) suggest in their work. In a short ablation study (see Appendix A), we show that our chosen disambiguation architecture works with similar performance as a simple binary classifier.

[1] We did some initial studies with BIOS and BIOES encodings, but the performance was not satisfying.

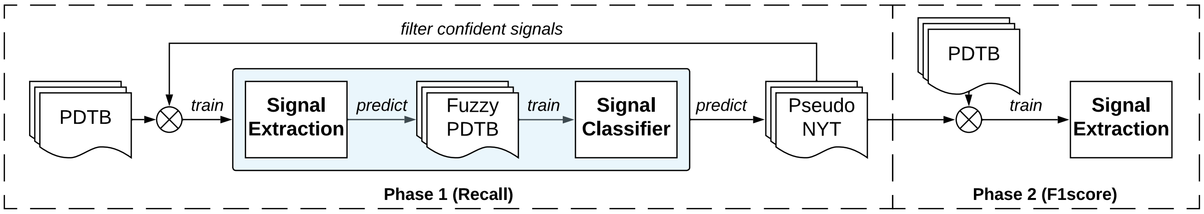

Figure 1: Overview of the learning process: Phase 1 refers to the cyclic self-supervised learning procedure (alternates between labeling and candidate discrimination). Phase 2 concludes the final training on combined data.

Our self-learning approach consists of two phases (compare Figure 1). In **Phase 1 (Recall)**, we optimize our signal labeling model (**Signal Extraction**) with respect to high recall. Therefore, we lower the weight for the **None** class, which is the dominating class label, and thus implicitly reinforce a higher focus on the other class labels. The resulting increase in the recall value simultaneously leads to a reduction in the precision value of the model. We first extract signal candidates from the PDTB which results in a fuzzy version, in order to train a second model for signal disambiguation (**Signal Classifier**). In the final step of a single iteration, we extract signals from a different corpus (here NYT see Section 4.1) and filter these signals, based on the confidence score, before we use the confident paragraph samples together with the original training data. Confident paragraphs are defined based on the individual signal confidence, such that signals are removed if the confidence is below a *relation threshold* τ_{rel} and the remaining signals' minimum is higher than a *paragraph threshold* τ_{par}. In **Phase 2 (F1Score)**, we use the finally extracted confident paragraphs and train a new model on signal extraction, but this time all class labels are equally weighted and the model is optimized for F1 score.

4 Experiments

4.1 Experimental Settings

The Unlabeled Corpus Most, but not all (Webber, 2009), documents of the PDTB are news articles. To learn about alternative lexicalizations from a different corpus, as there is relatively little annotated data currently available, we choose another news wire corpus, under the assumption of an easier adaption of a similar domain compared to other genres. *The New York Times Annotated Corpus* [2], referred to as NYT, contains about 1.8 million documents published by the New York Times between 1987 and 2007. For our work, we use a random subset of documents, 200 per month from the years 2000–2002, sampled only once before the experiments. The reduced corpus is due to computational feasibility, the years were selected randomly. We selected NYT to complement the PDTB training data because much more data is available and it has similar quality and structure of articles as in the Wallstreet Journal corpus, which is used for the PDTB. For example, we decided against the CNN/DM corpus used in a different study (Kobayashi et al., 2021) because of the largely absent paragraph structure.

Hyper-Parameter Settings For data preparation, we split 10% of documents from the PDTB corpus for testing purposes. While we use varying test splits for the general evaluation of the architecture, also to compare to previous work, we use the same test split for the evaluation of the self-supervised setting. After creating a separate test set, in each run, we split another 10% of the remaining training documents for validation. To increase the reproducibility of our experiments, we use the same validation splits for each model run, e.g. we have the same 3 and 5 splits for model ensembles and evaluations, respectively. For both types of models, signal labeling and sense classification, the batch size is 32. We train for at most 10 epochs and stop after 3 epochs without any improvement. For optimization, we chose Adam with decoupled weight decay (Loshchilov and Hutter, 2019) and an initial learning rate of 1e−4 that is reduced linearly over the maximum training epochs. As we observed overfitting with a too-small dropout rate, we set it to 0.4 for both models. For embedding paragraphs, we chose the base architecture of RoBERTa (Liu et al., 2019) that has shown good performance on several other tasks related to discourse processing (Long and Webber, 2022; Koto et al., 2021; Guz et al., 2020). We fix all but the last two layers

[2] https://doi.org/10.35111/77ba-9x74

for signal labeling. For the disambiguation model, we extract all hidden units from the model and propagate them to our classifier. As the input size of RoBERTa is limited, we truncate the tokenized paragraph. Only less than 1% of the paragraphs are affected by this truncation. For signal classification, we remove training examples where a signal occurs after the limit.

During the adaption phase, we focus on the recognition of alternative lexicalizations rather than whether predictions are correct or not, as we later train an additional model that filters wrong predictions. We identify two crucial hyper-parameters: First, we examine changing the **majority class weight** (None class) for the cross-entropy loss. Second, we study the influence of **negative samples** on the training progress. In agreement with our experiments (for details see Appendix B), we chose 0.01 for the majority class weight as the next step's small increase in recall did not justify the higher decrease in precision. Further, our results indicate that there is no advantage in reducing the number of negative samples.

For both phases, we set the relation threshold τ_{rel} to 0.33 as we measured a good balance of true and false predictions on the PDTB data. For the paragraph threshold τ_{par} we use a value of 0.7 during training, as we focus on optimizing the recognition rate (recall) of the extraction model in this phase. In the second phase, we study varying thresholds ranging from 0.4 to 0.9 for minimal paragraph relation confidence.

4.2 Experimental Results

First, we evaluate our novel architecture and compare its base performance with the initial work by Knaebel and Stede (2022). In their work, they measure the overlap within sentences containing an alternative lexicalization. We, therefore, re-run their neural labeling model and use the same evaluation metrics (exact–match) as for this paper. Results are averaged over 10 random splits and presented as mean (M) and standard deviation (SD). In our evaluation under similar conditions, the baseline (M=34.07% F1, SD=6.09) is clearly outperformed by our introduced model (M=45.48% F1, SD=5.08). We also study the performance of ensembles as used in our self-learning setting and simply combine the output probabilities of three random models. The performance further improves (M=51.68% F1, SD=3.28) and we observe a decreasing standard deviation.

Results of our final experiments are shown in Figure 2 and in more detail in Appendix D. We compare the baseline trained on the original PDTB dataset with models of varying paragraph thresholds τ_{par} (0.4 to 0.9) that incorporate data from the NYT corpus into their training data. We utilize partial matching as introduced by Xue et al. (2016), and define the matching overlap based on the F1 score of two connecting phrases. *Partial-Match* and *Exact-Match* refer to 70% and 90% F1 matching thresholds, respectively. For example, our model recognizes two of three words of the signal *greatly expanding collaboration* correctly, resulting in 0.66 recall, 1.0 precision, and thus 0.83 F1, this signal would count as partially matched but not exactly. All Experiments run on the same test set, with varying training and validation splits, 5 repetitions each. Interestingly, all models perform best at a τ_{par} of 0.6, which is in accordance with the threshold suggested by Nishida and Matsumoto (2022). Our model (M=47.38% F1, SD=1.22) with all unlabeled data and $\tau = 0.6$ improves the baseline (M=42.95% F1, SD=2.52) by more than 4% F1 score on exact match.

4.3 Analysis of Selected Cases

In this section, we would like to show some selected signal examples that we noticed while reviewing the results. First, we look at the predictions of our recall-optimized signal extraction model (without filtering predictions by our second classifier) within the PDTB training data. This model has repeatedly recognized phrases (*after*, *and*, *on the other hand*, *at the same time*, *further*, *if*, *because [of]*, among others) as alternative lexicalizations although, in terms of their surface form, they should rather belong to the group of explicit connectors. We assume some of these phrases are only partially recognized alternative phrases e.g. signals in which the referential expression is missing *after this situation* and *because of that event*. We also identify cases where individual parts of the signal belong to explicit connectives, while their conjunction is rather considered as an alternative lexicalization, e.g. *since* and *then*. Despite a large number of possible explicit signals, most of the confused signals are filtered in the second step and are therefore not considered signals at all. Interestingly, we noticed that the model identifies a few signals at the beginning of a paragraph, similar as discussed by Prasad

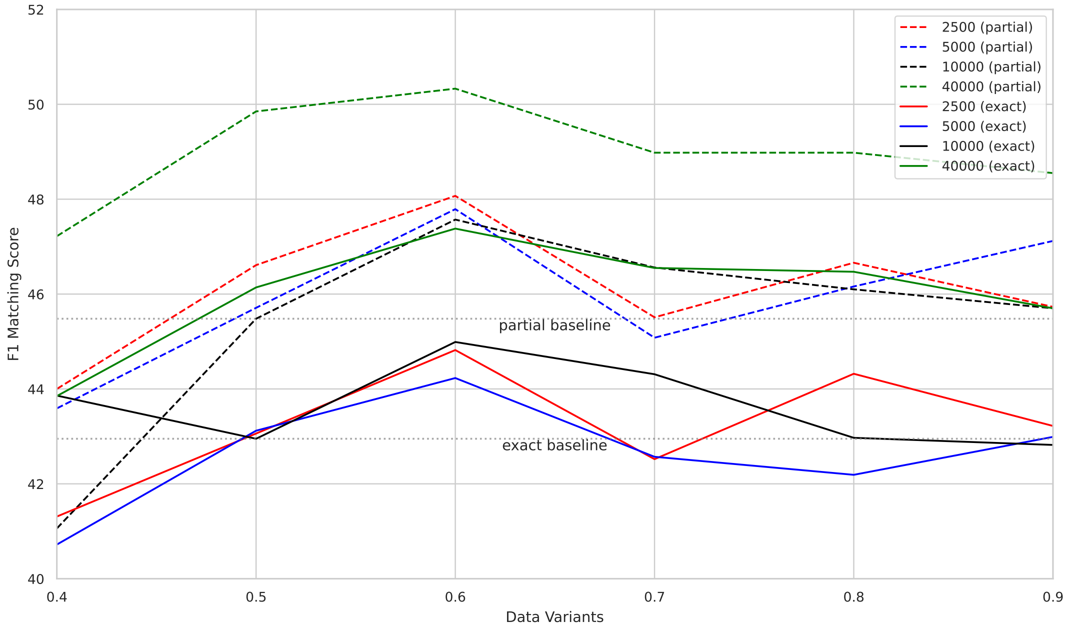

Figure 2: Final model evaluation: Comparison of baseline trained on the original dataset (horizontal dotted) and final models trained on data including NYT corpus with varying paragraph threshold τ_{par} (0.4 to 0.9) during prediction phase. All experiments run on the same test set, with varying training and validation splits, 5 repetitions. Evaluation is done using partial (dashed lines) and exact (straight lines) matching as explained in Section 4.2.

et al. (2017), that are per definition not included in the PDTB annotations, e.g. *That explains why*, *To illustrate*, *All this illustrates that*, and *What's more*. Besides different variations of gerunds, we found phrases such as *at the most*, *that may mean*, and *even that* that are likely being used as discourse signals without checking their context. The integration of the model's output in future annotation processes may be beneficial in identifying more discourse signals.

Next, we examine the predicted alternative lexicalizations in the NYT data. Here, we found quite a few change verbs, e.g. *dimishing*, *bolstering*, *stimulating*, *absorbing*, *contributing* and *negating*, that never occurred in the training data. There is about the same number (20 each) of variants of alternative lexicalizations containing the word *reason* in both data sets which have no overlap with the respective other data set, e.g. *the reason is probably that*, *the reasons for that finding*, and *that alone is reason for* where identified in NYT but not in PDTB. We further found, that our model tends to predict shorter signals (average length of 9) compared with the PDTB training dataset (average length of 13). The longest extracted signals with respect to token counts are *one reason for the cooperative ads is that*, *the overhaul was spurred in part by*, and *and that might partly explain why*.

5 Discussion and Conclusions

We developed a new paragraph-based architecture to extract alternatively–lexicalized discourse signals and presented state-of-the-art performance. Initial experiments on incorporating non-annotated data showed a further increase in performance.

Size seems to matter for this learning too, as this principle often holds for deep learning models. Although the gaps are rather small for up to 10,000 sampled documents, we think the distance for the largest set of documents is very clear. Due to time and computation constraints, we could not identify an upper performance bound yet.

We notice throughout our signal extraction experiments a confusion between alternative lexicalizations and explicit connectives. We assume the model to have problems clearly understanding their difference, as both kinds of phrases signal discourse relations. Filtering the connecting phrases as we have done seems unavoidable. Contrary to this, however, it seems worthwhile to soften the boundaries between these two categories and develop models that combine both types. This is not trivial due to the differences between both signal types (explicit signals are usually shorter recurring phrases with higher frequencies; AltLex signals tend to be longer phrases with more variance).

Limitations

Although the new architecture works well on PDTB-like structured data, we are often challenged with texts without clear paragraph structure. This would make it either necessary to pre-process texts and split sentences into semantically closed paragraphs such that our proposed model takes advantage of the surrounding context, or develop a new sentence-based model which was not successful in previous work.

Limiting the model to predict only continuous alternative lexicalizations does not highly affect results on the PDTB, but might have a more considerable impact on other text genres, e.g. speeches and debates. This would require the use of a more complex signal encoding as mentioned in Section 3.1.

Acknowledgements

This research has been supported by the German Research Foundation (DFG) with grant number 455911521, project "LARGA" in SPP "RATIO". I would like to thank Manfred Stede and Freya Hewett as well as the anonymous reviewers for their valuable feedback.

References

Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. Multi-view and multi-task training of RST discourse parsers. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1903–1913, Osaka, Japan. The COLING 2016 Organizing Committee.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Chien-Lung Chou, Chia-Hui Chang, and Shin-Yi Wu. 2014. Semi-supervised sequence labeling for named entity extraction based on tri-training: Case study on Chinese person name extraction. In *Proceedings of the Third Workshop on Semantic Web and Information Extraction*, pages 33–40, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70.

Grigorii Guz, Patrick Huber, and Giuseppe Carenini. 2020. Unleashing the power of neural discourse parsers - a context and structure aware approach using large scale pretraining. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3794–3805, Barcelona, Spain (Online). International Committee on Computational Linguistics.

René Knaebel and Manfred Stede. 2020a. Contextualized embeddings for connective disambiguation in shallow discourse parsing. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 65–75, Online. Association for Computational Linguistics.

René Knaebel and Manfred Stede. 2020b. Semi-supervised tri-training for explicit discourse argument expansion. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1103–1109, Marseille, France. European Language Resources Association.

René Knaebel and Manfred Stede. 2022. Towards identifying alternative-lexicalization signals of discourse relations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 837–850, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2021. Improving neural RST parsing model with silver agreement subtrees. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1600–1612, Online. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Murathan Kurfalı. 2020. Labeling explicit discourse relations using pre-trained language models. In *Text, Speech, and Dialogue*, pages 79–86, Cham. Springer International Publishing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Wanqiu Long and Bonnie Webber. 2022. Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10704–10716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Reranking and self-training for parser adaptation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 337–344, Sydney, Australia. Association for Computational Linguistics.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

Rashmi Prasad, Katherine Forbes Riley, and Alan Lee. 2017. Towards full text shallow discourse relation annotation: Experiments with cross-paragraph implicit relations in the PDTB. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 7–16, Saarbrücken, Germany. Association for Computational Linguistics.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Realization of discourse relations by other means: Alternative lexicalizations. In *Coling 2010: Posters*, pages 1023–1031, Beijing, China. Coling 2010 Organizing Committee.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Magdaléna Rysová and Kateřina Rysová. 2015. Secondary connectives in the Prague dependency treebank. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 291–299, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Mark Steedman, Rebecca Hwa, Stephen Clark, Miles Osborne, Anoop Sarkar, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Example selection for bootstrapping statistical parsers. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 236–243.

Jun Suzuki and Hideki Isozaki. 2008. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *Proceedings of ACL-08: HLT*, pages 665–673, Columbus, Ohio. Association for Computational Linguistics.

Pavlína Synková, Magdaléna Rysová, Lucie Poláková, and Jiří Mírovský. 2017. Extracting a lexicon of discourse connectives in Czech from an annotated corpus. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 232–240. The National University (Phillippines).

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. 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*, pages 674–682, Suntec, Singapore. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Model	Precision	Recall	F1
Baseline	83.54 (5.99)	59.45 (12.34)	68.84 (9.68)
Coarse	86.32 (4.83)	51.78 (9.22)	64.36 (7.82)
Binary	85.18 (4.81)	56.81 (8.80)	67.80 (6.87)

Table 1: Ablation Study for the AltLex Candidate Classifier. Results show mean and standard deviation for 10 runs each.

A Candidate Disambiguation: Ablation Study

Discourse signal disambiguation is a fundamental step in our weakly-supervised learning cycle for improving the prediction quality of our signal extraction model. We intuitively followed previous work on signal-based sense classification (Knaebel and Stede, 2020a) with the assumption of better results learning multiple sense levels at once. (Long and Webber, 2022) Our ablation study in Table 1 shows, that contrary to our assumption the baseline and a binary classifier that is limited to predicting the discourse usage of a free connective phrase have similar performances. Removing the model's fine-sense classification drastically reduces the recall of identified signals but increases precision. This holds for the binary case, too. Further investigations are necessary to identify specific differences in these classifiers.

B Hyperparameters: Loss Weight and Negative Sampling

During the adaption phase, we focus on the recognition of alternative lexicalizations rather than whether predictions are correct or not, as we later train an additional model that filters wrong predictions. Therefore, we adjust the majority class weights (None class) of the cross-entropy loss. In Table 2a, we report macro averaged results for weights ranging from 1.0 (normal weight) to 0.001 (inverse occurrence weight). As expected, the results indicate an increase in average recall with a decrease in average precision at the same time. We chose 0.01 for the majority class weight as the next step's small increase in recall did not justify the higher decrease in precision.

We also study the influence of negative samples on the training progress. The results in Table 2b indicate no advantage of reducing negative samples for training data, as already mentioned so in the paper. However, in contrast, a broader study with varying test partitions showed an increase in recall while reducing the number of negative samples.

Weight	Precision	Recall	F1
1.0	41.63 (1.81)	32.75 (2.22)	36.63 (1.88)
0.5	33.93 (1.06)	39.49 (2.46)	36.43 (0.53)
0.1	21.41 (3.35)	51.52 (1.54)	30.03 (3.23)
0.01	8.13 (0.40)	61.39 (0.37)	14.35 (0.61)
0.001	3.59 (0.77)	63.34 (1.20)	6.78 (1.38)

(a) Weighting the majority class: None. '1.0' refers to normal training while '0.001' is close to the inverse of the class occurrences. By Reducing the None class weight, errors on remaining classes are stronger penalized, and thus the model parameters are optimized for recall.

ratio	Precision	Recall	F1
0.0	39.37 (1.29)	35.68 (1.83)	37.42 (1.44)
0.2	40.96 (2.54)	33.14 (0.24)	36.60 (0.95)
0.4	35.78 (1.15)	33.63 (1.59)	34.63 (0.74)
0.6	33.20 (1.89)	32.55 (1.33)	32.87 (1.61)
0.8	25.65 (1.51)	35.19 (1.50)	29.66 (1.43)
1.0	13.02 (0.40)	32.45 (4.37)	18.48 (0.36)

(b) Down-sampling paragraphs without alternative lexicalizations as a performance factor, range from no sampling at all to remove all negative samples.

Table 2: Experiments on hyper-parameter settings for optimizing recall during the first training phase.

τ	2500	5000	10000	40000
0.4	1973	3883	7751	123595
0.5	1423	2816	5690	90016
0.6	572	1110	2256	37472
0.7	308	605	1205	19805
0.8	148	282	607	10216
0.9	46	91	200	3495

Table 3: Number of training samples extracted from additional pseudo labeled corpus, per corpus sample size and per relation paragraph threshold.

C Numbers of Extracted Paragraphs

Table 3 summarizes the number of training samples that were extracted from a given corpus sample (limited by the number of documents) and a corresponding relation paragraph threshold that needs to be satisfied for positive training samples.

D Full Final Results

Table 4 summarizes our final experiments' results in full detail. Partial-Match and Exact-Match refer to 70% and 90% overlap, respectively. In contrast to the evaluation with previous work, for this evaluation, we split test data only once at the very beginning and stay with it throughout the evaluation. Results are averaged over different validation splits, though.

Model	Partial-Match			Exact-Match		
	Precision	Recall	F1	Precision	Recall	F1
Base	41.08 (4.95)	51.78 (3.58)	45.48 (2.26)	38.82 (5.10)	48.84 (2.90)	42.95 (2.52)
0.4	36.50 (1.50)	55.50 (1.74)	44.00 (0.90)	34.28 (1.62)	52.09 (1.24)	41.31 (0.98)
0.5	38.72 (3.44)	59.22 (3.01)	46.61 (1.50)	35.76 (3.01)	54.73 (3.05)	43.06 (1.29)
0.6	43.02 (1.49)	54.73 (3.27)	48.07 (0.77)	40.12 (1.84)	51.01 (3.00)	44.82 (1.18)
0.7	40.70 (5.78)	53.02 (4.48)	45.51 (1.86)	38.09 (6.12)	49.46 (3.51)	42.52 (2.36)
0.8	42.20 (2.87)	52.40 (2.23)	46.66 (1.67)	40.09 (2.79)	49.77 (1.79)	44.32 (1.52)
0.9	40.41 (3.33)	53.18 (3.49)	45.73 (1.83)	38.20 (3.51)	50.23 (3.19)	43.22 (2.06)

(a) NYT corpus (2500 documents).

Model	Partial-Match			Exact-Match		
	Precision	Recall	F1	Precision	Recall	F1
Base	41.08 (4.95)	51.78 (3.58)	45.48 (2.26)	38.82 (5.10)	48.84 (2.90)	42.95 (2.52)
0.4	34.97 (1.85)	58.29 (3.58)	43.59 (0.79)	32.69 (2.23)	54.42 (3.19)	40.72 (1.51)
0.5	38.45 (2.06)	56.74 (3.71)	45.71 (1.11)	36.29 (2.22)	53.49 (2.73)	43.12 (1.00)
0.6	40.51 (1.38)	58.29 (3.08)	47.79 (1.89)	37.49 (1.15)	53.95 (2.76)	44.23 (1.62)
0.7	38.78 (3.43)	54.57 (4.06)	45.08 (1.68)	36.66 (3.66)	51.47 (3.12)	42.57 (1.92)
0.8	39.58 (1.82)	55.50 (1.74)	46.16 (1.09)	36.19 (2.24)	50.70 (1.26)	42.19 (1.59)
0.9	42.54 (3.72)	53.64 (4.29)	47.12 (0.67)	38.77 (3.01)	48.99 (4.64)	42.99 (0.72)

(b) NYT corpus (5000 documents).

Model	Partial-Match			Exact-Match		
	Precision	Recall	F1	Precision	Recall	F1
Base	41.08 (4.95)	51.78 (3.58)	45.48 (2.26)	38.82 (5.10)	48.84 (2.90)	42.95 (2.52)
0.4	32.27 (1.87)	56.90 (3.42)	41.06 (0.82)	34.47 (2.06)	60.78 (3.56)	43.86 (0.90)
0.5	41.08 (4.95)	51.78 (3.58)	45.48 (2.26)	38.82 (5.10)	48.84 (2.90)	42.95 (2.52)
0.6	41.26 (2.75)	56.43 (2.75)	47.57 (1.99)	39.03 (3.03)	53.33 (2.37)	44.99 (2.24)
0.7	40.75 (1.83)	54.42 (2.57)	46.56 (1.51)	38.79 (1.80)	51.78 (2.05)	44.31 (1.31)
0.8	41.42 (3.25)	52.40 (2.48)	46.10 (1.23)	38.61 (3.10)	48.84 (2.25)	42.97 (1.32)
0.9	40.24 (3.44)	53.18 (2.11)	45.70 (2.21)	37.74 (4.16)	49.77 (2.37)	42.82 (3.11)

(c) NYT corpus (10000 documents).

Model	Partial-Match			Exact-Match		
	Precision	Recall	F1	Precision	Recall	F1
Base	41.08 (4.95)	51.78 (3.58)	45.48 (2.26)	38.82 (5.10)	48.84 (2.90)	42.95 (2.52)
0.4	38.29 (1.94)	**62.02** (4.33)	47.22 (1.42)	35.58 (2.18)	**57.52** (2.75)	43.85 (0.98)
0.5	42.51 (1.30)	60.31 (1.42)	49.85 (0.99)	39.35 (1.62)	55.81 (1.77)	46.14 (1.49)
0.6	44.13 (2.17)	58.60 (1.60)	**50.33** (1.80)	41.54 (1.53)	55.19 (1.42)	**47.38** (1.22)
0.7	43.68 (1.96)	55.81 (0.49)	48.98 (1.17)	41.52 (2.49)	53.02 (1.05)	46.55 (1.87)
0.8	**44.60** (3.03)	54.42 (1.14)	48.98 (2.05)	**42.32** (3.09)	51.63 (1.67)	46.47 (2.30)
0.9	43.54 (3.95)	55.19 (1.24)	48.55 (2.14)	41.00 (4.01)	51.94 (1.47)	45.70 (2.46)

(d) NYT corpus (40000 documents).

Table 4: Full results, partial and exact matching, of final model with varying paragraph threshold (0.4 to 0.9) trained on data including NYT corpus. All experiments run on the same test set, with varying training and validation splits, results are averaged over 5 repetitions.