Using Masked Language Model Probabilities of Connectives for Stance Detection in English Discourse

Regina Stodden[1], **Laura Kallmeyer**[1], **Lea Kawaletz**[2] and **Heidrun Dorgeloh**[2]
[1] Institute of Linguistics; [2] Institute of English and American Studies
Heinrich Heine University
Düsseldorf, Germany
{firstname.secondname}@hhu.de

Abstract

This paper introduces an approach which operationalizes the role of discourse connectives for detecting argument stance. Specifically, the study investigates the utility of masked language model probabilities of discourse connectives inserted between a claim and a premise that supports or attacks it. The research focuses on a range of connectives known to signal support or attack, such as *because, but, so,* or *although*. By employing a LightGBM classifier, the study reveals promising results in stance detection in English discourse. While the proposed system does not aim to outperform state-of-the-art architectures, the classification accuracy is surprisingly high, highlighting the potential of these features to enhance argument mining tasks, including stance detection.

1 Introduction

The task this paper addresses is argument stance detection in English discourse. More concretely, based on the definition of *argument* following established terminology (Stab and Gurevych, 2017; Stede and Schneider, 2018), where an argument consists of a *claim*, a controversial statement, and a *premise*, a statement supporting or attacking the claim, we want to automatically decide whether the premise supports (label: 1) or attacks (label: 0) the claim. This task has been modeled in a number of approaches already (Schiller et al., 2021; Hardalov et al., 2021). In contrast to these approaches, we aim at operationalizing the role of connectives with the following simple idea: We insert one-word connectives, i.e., linking words such as *because, but, so,* or *although*, between the claim and the candidate premise and use a language model (LM) to quantify acceptability. Connectives include coordinators (such as *and*, or *but*), subordinators (such as *because*, or *while*), as well as linking adverbs (such as *therefore*, or *however*; Dorgeloh and Wanner 2022). They can express support, attack, or other types of relations. The underlying hypothesis is that features obtained from an LM's probability for inserting certain connectives between a claim and premise can improve stance detection. Put differently, our research question is whether we can verify whether a premise is a support for or an attack against a given claim based on explicit discourse connectives. We show that using probabilities of connectives as features, we obtain a significant improvement in stance detection compared to a majority and a random baseline. This indicates that, although we do not aim at a competitive argument mining system in this paper, integrating these features into argument mining has the potential to improve existing approaches. We use English data but we assume that a similar approach should also work for other languages.[1]

2 Motivation and Related Work

The expression of stance is linked closely to argumentative structures in discourse since arguments by definition involve stance, and stance markers are known to facilitate the processing of argumentative relations (Stein and Wachsmuth, 2019; Wei et al., 2021). Besides a variety of other stance markers (Gray and Biber, 2014), connectives play a crucial role in that respect. Work on various languages has shown that the discourse function of connectives is closely related to that of other linguistic elements expressing stance or subjectivity in their role for argumentative discourse. In particular, there seems to be a "division of labor," where the presence of stance markers makes an explicit connective less expected while fewer stance markers make the use of specific connectives more likely (Wei et al. 2020). Such a trade-off between connectives and other cues for stance suggests that markers of one kind may be omitted if there are cues in the context that make the information of those markers

[1]The code and results are available at https://github.com/rstodden/stance-detection.

Proceedings of the 10th Workshop on Argument Mining, pages 11–18
December 7, 2023.

already predictable (Uniform Information Density Hypothesis; Torabi Asr and Demberg 2015), which motivates here our expectation that discourse connectives also mark argument stance.

Masked LMs (MLMs), e.g., BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019), are bidirectional encoders which are mostly trained on massive data to solve the task of *language modeling*. The intention of language modeling is similar to a cloze test; the model is trained on extensive unlabeled data, wherein random tokens (at any position of a sequence) are masked, enabling the model to learn how to predict them (Devlin et al., 2019). The pre-trained MLMs return probabilities for any word of the vocabulary at the position of the masked token; the higher the probability the more suitable the word in the sequence. In recent years, MLMs are also often used for stance detection.[2] Following Schiller et al. (2021), the current state-of-the-art model (called MT-DN$_{MDL}$) across multiple stance detection datasets is a BERT model (bert-large-uncased with an additional classification layer; Devlin et al. 2019), initially fine-tuned on the GLUE benchmark (Wang et al., 2018) and subsequently fine-tuned concurrently on several stance detection datasets . In contrast to MT-DN$_{MDL}$ and related models, in our approach we do not predict the stance based on the weights of an MLM but make use of the knowledge of MLMs with respect to connectives as stance markers.

Methodologically, the present study builds on existing approaches which tackle the problem of classifying implicit discourse relations by using masked LMs to *explicitate* the relations. Specifically, the models predict how likely a given connective is in sentence pairs without an overtly expressed discourse relation. For example, Kishimoto et al. (2020) experiment with additionally pre-training and fine-tuning MLMs on texts with masked connectives (called *connective prediction task*), finding that only the first technique provides gain. Kurfalı and Östling (2021) use a pipeline approach to classify unlabeled, implicit discourse relations, where explicit data – a set of 65 candidate connectives – is concatenated with two sequences and then fed into an explicit discourse relation classifier. Recently, Zhou et al. (2022) have tackled the problem by using a prompt learning method. Given a template that arises from natural language use (e.g. 'Arg1: *Arg1*. Arg2: *Arg2*. The conjunction between Arg1 and Arg2 is <mask>.'), they select the most frequent and least ambiguous predicted connective as the answer word to replace the mask token. We do not use prompting or causal LMs as we are interested in the probabilities of the connectives. Masked LMs, in contrast to generative LMs, are capable of giving the probabilities of a word at any position of a sequence based on the left and right context (and not only at the end of a sequence). To the best of our knowledge, our work is the first approach to use probabilities of discourse connectives of masked LMs as features and to combine them with stance detection.

3 Methods

Our method comprises four components:

1. the concatenation of claims and premises with a masked token (see subsection 3.2),
2. an LM that estimates the likelihood of a given connective in the concatenated sequence (see subsection 3.1 and subsection 3.3),
3. a feature vector which comprises all the probabilities of the connectives (see subsection 3.3),
4. and a binary classifier which, based on the feature vector, learns whether the premise supports or attacks the claim (see subsection 3.4).

We hypothesize that the LMs have learned argumentative structures and the usage of connectives. Therefore, we anticipate that the model will assign higher probabilities to support connectives and lower probabilities to attack connectives for support premises, and vice versa for attacks. For example, in Example 1, the premise attacks the claim, and we expect lower probabilities for support connectives like *because* or *since* as they would render the argument incoherent. For attack connectives like *but* or *although* we expect higher probabilities as they are in line with the attack relation.

(1) [Masking should be mandated]$_{\texttt{C}}$ `[MASK]` [it infringes on personal freedoms.]$_{\texttt{P}}$

3.1 Connectives

We selected connectives from DimLex-Eng (Das et al., 2018), a lexicon of discourse markers which contains 100 connectives from the Penn Discourse Treebank (PDTB; Prasad et al. 2008) plus 42 from RST-SC (Das and Taboada, 2018), all annotated with discourse relations. Out of all 79 single-token connectives, we selected

[2]For an overview of existing stance detection datasets and approaches see Schiller et al. (2021); Hardalov et al. (2021).

those with relevant PDTB relations[3]: For the support relation we chose connectives marked with `Contingency.Cause.Result` or `Contingency.Cause.Reason` (n=18, e.g. *therefore*, *because*), since relations in the contingency class "involve an implication relation, and hence can be classified as causal" (Sanders et al., 2021, 21). For the attack relation we chose `Comparison.Contrast` and `Expansion.Alternative.Disjunctive` (n=30, e.g. *but*, *however*), as they correlate with the attack relations of *undercut* and *rebut* (Hewett et al. 2019). Finally, some connectives were excluded as the LMs tokenized them into subwords (e.g., *however*: *how* and *ever*).[4] Table 1 summarizes the resulting 12 support-indicating and 18 attack-indicating connectives for which probabilities could be extracted.[5] Six connectives are labeled with the relations of both groups.

For more information on the connectives, we calculated how often a connective is tagged with the chosen PDTB relations divided by the number of all occurrences of the connective in DimLex-Eng. Based on this percentage, we grouped the connectives as follows: *Group 1*: all attack/support connectives (>0%, n=24), *Group 2*: not predominatly attack/support connectives (>34%, n=12), i.e., those which were used in up to 66% of occurrences in some other PDTB relation, and *Group 3*: predominantly attack/support connectives (>66%, n=5), i.e., those which were used in up to 34% of occurrences in some other PDTB relation.

attack				support			
conn.	**order**	**%**	**G**	**conn.**	**order**	**%**	**G**
unless	C-LW-P	98.95	1,2,3	for	C-LW-P	100.0	1,2,3
but	*C-LW-P*	*73.28*	*1,2,3*	so	P-LW-C	100.0	1,2,3
while	C-LW-P	52.50	1,2	because	C-LW-P	99.53	1,2,3
yet	P-LW-C	52.48	1,2	with	C-LW-P	60.00	1,2
still	P-LW-C	50.53	1,2	since	C-LW-P	52.17	1,2
although	C-LW-P	47.87	1,2	given	C-LW-P	33.33	1
though	C-LW-P	47.50	1,2	*as*	*C-LW-P*	*28.53*	*1*
rather	P-LW-C	23.53	1	*and*	*C-LW-P*	*2.17*	*1*
except	C-LW-P	10.00	1	*when*	*C-LW-P*	*2.02*	*1*
nor	C-LW-P	3.23	1	*then*	*C-LW-P*	*1.47*	*1*
instead	C-LW-P	2.68	1	*if*	*C-LW-P*	*0.08*	*1*
until	C-LW-P	1.85	1	*but*	*C-LW-P*	*0.03*	*1*
or	C-LW-P	1.02	1				
and	*C-LW-P*	*0.70*	*1*				
if	*C-LW-P*	*0.41*	*1*				
then	*C-LW-P*	*0.29*	*1*				
when	*C-LW-P*	*0.20*	*1*				
as	*C-LW-P*	*0.13*	*1*				

Table 1: Connectives with their order (*claim-connective-premise* or *premise-connective-claim*) and usage in PDTB as attack (left) or support (right). `G` shows the group of the connectives for the analysis. Connectives in italics are both attack as well as support.

3.2 Data & Preprocessing

In comparison to Hardalov et al. (2021) and Schiller et al. (2021), we reduce the selection of corpora to the following three corpora: `ibmcs` (Bar-Haim et al., 2017), `perspectrum` (Chen et al., 2019), and `argmin` (Stab et al., 2018).[6] All corpora (except `argmin`) have full sentences as claims (= topics) and have (balanced) binary stance labels.[7] For `argmin`, we changed the one-word topics to sentences (e.g., for topic *"cloning"*: *"cloning should be permitted."*).

During preprocessing, we remove any given punctuation mark at the end of the first argument component and lower-case the beginning of the second part. We then concatenate each pair of premise and claim with a masked token, e.g., "<mask>," that indicates the place for a potential connective. For every argument, we create the concatenation in the following two orders, because not all connectives require the same order of claim and premise (see Table 1): i) claim - masked token - premise (order C-LW-P), or ii) premise - masked token - claim (order P-LW-C). Some examples of the concatenated sequences are provided in the Appendix, Table 5. We do not tokenize the data or do any other preprocessing beyond what has already been mentioned (or is provided in the original corpus).

3.3 Feature Extraction

We then use these concatenated sequences as input for a masked LM, e.g., BERT (Devlin et al., 2019). As output, the LM returns word-probability pairs, where words with higher probabilities are more likely to be a suitable fit within the sequence.

We use the pipeline `fill-mask` of the Python package `transformers` (Wolf et al., 2020) to extract the probabilities of the connectives for

[3]We excluded all multi-token connectives as the applied fill-mask pipeline can predict only one token at a time.

[4]We do not employ Huggingface's fallback strategy, which is using subwords instead of the full word, as it could result in overly general word fragments (e.g., *how* for *however*).

[5]We also extracted all connectives which do not belong to any of the groups (n=13), henceforth called *other*. For DistilBERT and BERT, probabilities of more connectives could be extracted. However, we found out that using the probabilities of more connectives (n=42) of both LMs as features could not outperform using fewer connectives of RoBERTa or XLM-RoBERTa. Hence, we only report results on the reduced connective set (n=24) for all LMs.

[6]An overview of the datasets' meta data can be found in Table 1 and 2 of Hardalov et al. (2021).

[7]For our experiments, we used the original train, validation, and test splits provided by the authors of the datasets.

the following large LMs: i) *DistilBERT-base-uncased* (Sanh et al., 2019), ii) *BERT-base-uncased & -large* (Devlin et al., 2019), iii) *RoBERTa-base & -large* (Liu et al., 2019), as well as iv) *xlm-RoBERTa-base & -large* (Conneau et al., 2020).

The probabilities of either one of those LMs or of all LMs were then used as features for a classifier.[8] The LMs were not explicitly trained on argumentative data or structures and they were not fine-tuned on any other data or task; rather, we use them in their original form as provided on HuggingFace (Wolf et al., 2020).

3.4 Classifier

To find the best classifier and its best parameters for stance detection on all three datasets, we built up a search space of parameters[9] and applied methods of the `optuna` package (Akiba et al., 2019) to find the best hyperparameter combination for each validation set. Based on the best parameter combination for all probabilities of all LMs with all attack and support connectives, we averaged the parameters per validation set. The resulting parameters were then used for all experiments on the test sets. LightGBM turned out to be the best classifier out of six classifiers[10], hence, we are reporting only the results with LightGBM using the best hyperparameter setting (see Appendix A).

3.5 Evaluation

For the evaluation protocol, we mostly follow Schiller et al. (2021); Hardalov et al. (2021): We evaluate our approaches by calculating the macro F1-Score, and we report a majority baseline (always returns the most frequent label) and a random baseline (randomly returns one label of the two labels). As further comparison, we also report results of four state-of-the-art models (SOTA): i) BERT-large with a classification head (BERT$_{SDL}$), ii) BERT fine-tuned on GLUE benchmark with a classification head (MT-DNN$_{SDL}$), iii) MT-DNN$_{SDL}$ additionally trained on ten stance detection data sets (MT-DNN$_{MDL}$; Schiller et al. 2021), and iv) RoBERTa-base with domain expert functions and a classification head (MoLe; Hardalov et al. 2021).

[8] An example of probabilities for given sequences is provided in the Appendix, Table 5.

[9] For the entire search space per classifier see the code.

[10] We have also experimented with the following classifiers and search spaces for them:i) a support vector machine, ii) a decision tree classifier, iii) a random forest classifier, iv) a neural multi-layer perceptron, and v) a XGBoost classifier.

4 Results

We first validated our main assumption by measuring Spearman's correlation coefficient ρ between the probabilities of the connectives and the stance per each sample of each dataset. Appendix B summarizes all correlations and significance levels. For all three datasets, we found that the probabilities of nearly all connectives significantly correlate with stance (p-level at least < 0.1; all except *with*, *if*, and *when*). As expected, the probabilities of the attack connectives show a negative correlation, whereas those of the support connectives show a positive correlation, and the ambiguous connectives show a mixed picture. However, most correlations are weak (i.e., $\rho < 0.3$) except for five moderate (i.e., $0.3 \leq \rho < 0.5$; *except*, *unless*, *until*, *yet*, and three strong ones (i.e., $\rho \geq 0.5$; *although*, *though*, *but*). To sum up, our assumption was validated across all three datasets. Therefore, we can now turn to our results on stance detection based on the connectives' probabilities.

All our models using all connectives (Group 1) can outperform the two baselines. The best model with all probabilities (Group 1) of only one LM is RoBERTa-large (see bold row in the third part of Table 2). As expected, DistilBERT achieves the worst results compared to all other LMs, and all large versions outperform their base versions. We can infer that the larger the model and the more data the model was trained on, the more knowledge it has about connectives and, therefore, the more valuable the connective features are for stance detection and, hence, the higher the macro F1-Score. However, the multi-lingual data on which xlm-RoBERTa is trained seems to reduce the score, which might be due to its larger vocabulary size and less distinct probabilities for the connectives. Further analysis is required to justify this finding. Overall, combining the probabilities of all 24 connectives (Group 1) of all LMs achieves a higher macro F1-Score than using the Group 1 probabilities of only one LM (see bold row in the last part of Table 2). This model outperforms all SOTA models on `argmin` and is on par with the SOTA model on the other two datasets. Comparing all models based on BERT-large (i.e., BERT$_{SDL}$, MT-DNN$_{SDL}$, MT-DNN$_{MDL}$, and our BERT-large), our model achieves similar scores as the other models on the `argmin` dataset, although it classifies just on the probabilities of 24 connectives of neither fine-tuned nor otherwise preprocessed LMs.

Further, we analyzed the ablation of some ambiguous connectives (see results of Group 2), e.g., *and* or *when*, and not predominant connectives, e.g., *instead* or *given* (see results of Group 3).

As can be seen in the last six lines of Table 2 (or also for all other LMs in the Appendix, Table 6), the ablations reduce the scores. The more support and attack connectives (or features), the better the result. It can be argued that not only distinctive connectives, such as *because* or *yet*, are helpful for stance detection, but also the presence of other connectives. Yet, adding additionally the probabilities of all *other* connectives (n=12), slightly reduces the F1-Score on `argmin` and `ibmcs` (see last row in Table 2), whereas it increases the score on `perspectrum`. Hence, the selection of the connectives is also important. For example, replacing the 24 support and attack connectives by 24 randomly chosen connectives (12 *other* and 12 randomly chosen support or attack connectives) the score drops on average of 5 runs. Further, including only the probabilities of the *other* connectives (n=12) reduces the score even more.

Also, the combination of attack and support connectives seems to be helpful for stance detection (see Appendix C). For all datasets, the F1-Score drops when removing support connectives (by less than 0.01 points) as well as, more noticeably, when removing attack connectives (between 0.01 and 0.35 points). When using only connectives which are in both lists (n=6), the score even drops by one more 0.01 point. This effect might be due to the decreasing number of features, as the analysis of the connectives of Group 3 with the same number of features (i.e., connectives most often used for attack or support, n=5) also show a clear drop in performance. An additional observation is that some connectives (e.g., *and*, *when*) appear in both groups, indicating that their interpretation as support or attack is inferred. This highlights that the role of connectives in signaling stance does not necessarily demand the explicit expression of the semantics of the claim-premise relation.

5 Conclusion and Future Work

In this paper, we performed stance detection based only on the masked LM probabilities of discourse connectives that are assumed to indicate support or attack. The classifiers we trained on these features performed surprisingly well, given that the aim was not at all to develop a competitive argument mining system. From our results one can conclude that connectives, i.e. different kinds of linking words, can help to automatically verify if a premise is related to a given claim and, with that, also aid stance detection. Connectives should thus play an even more prominent role in argument mining.

In future work, we plan to also experiment with additional punctuation marks between the first part and the linking word. This is a promising avenue because some connectives occur more naturally at a sentence beginning and not between two clauses, e.g., *therefore*, or require a preceding comma, e.g., *but*. Furthermore, we plan to integrate features based on the MLM probabilities of connectives, as used in this paper, with state-of-the-art approaches to stance detection that use input embeddings representing the actual text of claim and premise. Finally, we will investigate whether additional preprocessing of the LMs in the form of fine-tuning on argumentative data or data with explicit connectives before extracting the MLM probabilities increases stance detection performance.

models	**argmin**	**ibmcs**	**perspectrum**
majority	0.3383	0.3406	0.3466
random	0.4998	0.4864	0.5011
BERT$_{SDL}$	0.6167	0.5347	0.8012
MT-DNN$_{SDL}$	0.6019	0.7066	0.8480
MT-DNN$_{MDL}$	0.6174	0.7772	0.8374
MoLe	**0.6373**	**0.7938**	**0.8527**
DistilBERT	0.5233	0.5499	0.6079
BERT-base	0.5718	0.5500	0.6442
BERT-large	0.6104	0.5810	0.6828
RoBERTa-base	0.6218	0.5961	0.6890
RoBERTa-large	**0.7204**	**0.7633**	**0.8274**
xlm-RoBERTa-base	0.5830	0.5456	0.6130
xlm-RoBERTa-large	0.6601	0.7247	0.7475
all-LMs (Group 1, n=24)	**0.7467**	**0.7885**	0.8314
all-LMs (Group 2, n=12)	0.7218	0.7638	0.8185
all-LMs (Group 3, n=5)	0.6861	0.7449	0.7897
all-LMs (other, n=12)	0.6792	0.6676	0.7539
all-LMs (random, n=24)	0.7286	0.7710	0.8286
all-LMs (all, n=36)	0.7423	0.7850	**0.8456**

Table 2: First part baselines, second SOTA, third own models per LM features (Group 1), and last combination of all feature groups of all LMs. Results of SOTA are copied from corresponding paper. F1 macro scores.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. We would also like to thank NVIDIA for access to GPUs, which enabled a fast calculation of the probabilities.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. 2018. Constructing a lexicon of English discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–365, Melbourne, Australia. Association for Computational Linguistics.

Debopam Das and Maite Taboada. 2018. RST Signalling Corpus: A Corpus of Signals of Coherence Relations. *Language Resources and Evaluation*, 52(1):149–184.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Heidrun Dorgeloh and Anja Wanner. 2022. *Discourse Syntax: English Grammar Beyond the Sentence*. Cambridge University Press.

Bethany Gray and Douglas Biber. 2014. Stance markers. In Karin Aijmer and Christoph Rühlemann, editors, *Corpus Pragmatics: A Handbook*, chapter 8, page 219–248. Cambridge University Press.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Murathan Kurfalı and Robert Östling. 2021. Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 1–10, Online. Association for Computational Linguistics.

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.

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).

Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2021. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 17(1):1–71.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *KI - Künstliche Intelligenz*, 35(3-4):329–341.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Manfred Stede and Jodi Schneider. 2018. *Argumentation Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publisher.

Benno Stein and Henning Wachsmuth, editors. 2019. *Proceedings of the 6th Workshop on Argument Mining*. Association for Computational Linguistics, Florence, Italy.

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform surprisal at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 118–128, London, UK. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yipu Wei, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. 2020. The use of perspective markers and connectives in expressing subjectivity: Evidence from collocational analyses. *Dialogue & Discourse*, 11:62–88.

Yipu Wei, Jacqueline Evers-Vermeul, Ted M. Sanders, and Willem M. Mak. 2021. The Role of Connectives and Stance Markers in the Processing of Subjective Causal Relations. *Discourse Processes*, 58(8):766–786.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3848–3858, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Hyperparameter of classifiers

Best hyperparameter: {"classifier": "LightGBM", "lambda_l1": 0.0001, "lambda_l2": 0.002, "num_leaves": 220, "feature_fraction": 0.9, "bagging_fraction": 0.8, "bagging_freq": 2}

B Correlation Connectives' Probabilities and Stance

	argmin	**ibmcs**	**perspectrum**
although	-0.24***	**-0.54*** **	**-0.53*** **
except	-0.26***	-0.49***	-0.44***
instead	-0.13***	-0.36***	-0.28***
nor	-0.15***	-0.23***	-0.24***
or	-0.04***	-0.12***	-0.10***
rather	-0.14***	-0.18***	-0.22***
still	-0.20***	-0.27***	-0.22***
though	-0.22***	**-0.53*** **	**-0.52*** **
unless	-0.2***	-0.35***	-0.36***
until	-0.18***	-0.34***	-0.32***
while	-0.11***	-0.37***	-0.21***
yet	-0.29***	-0.45***	-0.41***
because	+0.04***	+0.17***	+0.08***
for	+0.07***	+0.07***	+0.13***
given	+0.02*	+0.07**	+0.04***
since	+0.07***	+0.18***	+0.06***
so	+0.08***	+0.05*	+0.03***
with	+0.00	-0.13***	+0.01
and	+0.09***	-0.13***	+0.09***
as	+0.06***	+0.14***	+0.12***
but	-0.32***	**-0.54*** **	**-0.58*** **
if	-0.01	-0.09***	-0.08***
then	-0.02*	-0.21***	-0.12***
when	-0.02	-0.28***	-0.13***

Table 3: First block attacking connectives, second supporting connectives, and third which are classified as both. The asterisks indicate the level of significance (*: $p < 0.1$, **: $p < 0.05$, ***: $p < 0.01$). The bold face numbers indicate a strong, significant correlation ($\rho \geq 0.5$), underlining a moderate, significant correlation ($\rho \geq 0.3$) and the gray numbers are not significant.

C Results per Connective Type

	argmin	**ibmcs**	**perspectrum**
attack+support (n=24)	**0.7467**	**0.7885**	**0.8314**
attack (n=18)	0.7305	0.7872	0.8288
support (n=12)	0.7265	0.7531	0.8164
both (n=6)	0.7132	0.7513	0.8058

Table 4: Results per connective set for all LMs.

ID	stance	claim-connective-premise	because	but	premise-connective-claim	so	yet
Train-23	0	[Nuclear energy should be permitted]$_C$ [MASK] [it should be banned from Australia. If terrorists come they can target the power plant and it would kill heaps of people .]$_P$	0.000010 $<$	0.009026	[It should be banned from Australia. If terrorists come they can target the power plant and it would kill heaps of people]$_P$ [MASK] [nuclear energy should be permitted]$_C$	0.002439 $<$	0.00033800
Train-2874	1	[Nuclear energy should be permitted]$_C$ [MASK] [nuclear plants also provide stability to the electrical grid , as their output is constant and reliable .]$_P$	0.000584 $>$	0.000067	[Nuclear plants also provide stability to the electrical grid , as their output is constant and reliable]$_P$ [MASK] [Nuclear energy should be permitted .]$_C$	0.000018 $>$	0.00000037
Train-9125	0	[Cloning should be permitted]$_C$ [MASK] [when we consider cloning , we must not blindly overlook its negative implications .]$_P$	0.000005 $<$	0.002498	[When we consider cloning , we must not blindly overlook its negative implications]$_P$ [MASK] [cloning should be permitted .]$_C$	0.000014 $<$	0.00001765
Train-7226	1	[Cloning should be permitted]$_C$ [MASK] [a cloned child could actually enhance the family relationship for otherwise childless couples .]$_P$	0.000880 $>$	0.000026	[A cloned child could actually enhance the family relationship for otherwise childless couples]$_P$ [MASK] [cloning should be permitted .]$_C$	0.000061 $>$	0.00000006

Table 5: Cherry-picked examples of the argmin dataset including masking input and probabilities of connectives in both claim-premise orders. The $<$ and $>$ signs show the expected relation between the support and attack connectives in examples with positive and negative stance. The examples represent the opinions of the annotators and not necessarily those of the authors of this paper.

	Group 1 (n=24)			**Group 2 (n=12)**			**Group 3 (n=5)**		
	argmin	**ibmcs**	**perspectum**	**argmin**	**ibmcs**	**perspectum**	**argmin**	**ibmcs**	**perspectum**
DistilBERT	0.5233	0.5499	0.6079	0.5120	0.5373	0.5753	0.5006	0.5331	0.5690
BERT-base	0.5718	0.5500	0.6442	0.5448	0.5314	0.5939	0.5213	0.5316	0.5593
BERT-large	0.6104	0.5810	0.6828	0.5705	0.5898	0.6366	0.5610	0.5494	0.6154
RoBERTa-base	0.6218	0.5961	0.6890	0.6019	0.5842	0.6508	0.5757	0.5709	0.6152
RoBERTa-large	0.7204	0.7633	0.8274	0.7080	**0.7670**	0.8021	0.6683	0.7422	0.7677
xlm-RoBERTa-base	0.5830	0.5456	0.6130	0.5678	0.5473	0.5899	0.5455	0.5530	0.5608
xlm-RoBERTa-large	0.6601	0.7247	0.7475	0.6171	0.7082	0.7287	0.6070	0.6921	0.7149
all_LMs	**0.7467**	**0.7885**	**0.8314**	**0.7218**	0.7638	**0.8185**	**0.6861**	**0.7449**	**0.7897**

Table 6: Results per LM and feature set.