Discourse Relation Embeddings:
Representing the Relations between Discourse Segments in Social Media

Youngseo Son   **Vasudha Varadarajan**   **H. Andrew Schwartz**
Department of Computer Science, Stony Brook University
{yson,vvaradarajan,has}@cs.stonybrook.edu

Abstract

Discourse relations are typically modeled as a discrete class that characterizes the relation between segments of text (e.g. causal explanations, expansions). However, such predefined discrete classes limit the universe of potential relations and their nuanced differences. Adding higher-level semantic structure to modern contextual word embeddings, we propose representing discourse relations as points in high dimensional continuous space. However, unlike words, discourse relations often have no surface form (relations are *inbetween two segments*, often with no explicit word or phrase marker), presenting a challenge for existing embedding techniques. We present a novel method for automatically creating *discourse relation embeddings* (DiscRE), addressing the embedding challenge through a weakly supervised, multitask approach. Results show DiscRE representations obtain the best performance on Twitter discourse relation classification (macro $F1 = 0.76$) and social media causality prediction (from $F1 = .79$ to $.81$), performing beyond modern sentence and word transformers, and capturing novel nuanced relations (e.g. relations at the intersection of causal explanations and counterfactuals).

1 Introduction

Relations between discourse segments (i.e., phrases rooted by a main verb phrases or clauses) have mostly been studied as discrete classes; most notably Penn Discourse Treebank (PDTB) (Prasad et al., 2008) and Rhetorical Structure Theory Discourse Treebank (RST DT) (Carlson et al., 2001) contain 43 and 72 types of discourse relations respectively. At the same time, such work has taken place over newswire, the domain of both the PDTB and RST. With many different relation classes over sophisticated schema, annotation is non-trivial prohibiting extensive development in new domains (e.g., social media). Thus, progress in developing,
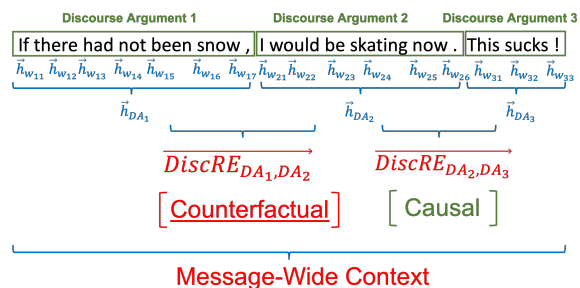

Figure 1: Our model DiscRE predicts relations of adjacent discourse arguments based on other text spans of the whole message as context. By learning and embedding fine-grained properties of discourse relation with the posteriors from PDTB into a continuous vector space, DiscRE may learn existing discourse relation targets like 'causal' relations, but also new latent discourse relations such as 'counterfactual' relations.

training and evaluating discourse relation identifiers has happened over *discrete-class* models with *labeled newswire* corpora (Pitler et al., 2009; Park and Cardie, 2012; Ji and Eisenstein, 2014; Lin et al., 2014; Popa et al., 2019).

To address this challenge and enable expansion of discourse work to social media, we propose a weakly supervised learning method which does not require any explicit labels. Instead, it adds a semantic structure that can effectively capture various types of discourse relations, even in other domains leveraging a multitask learning method called "Discourse Relation Embeddings (DiscRE)". Our DiscRE model represents discourse relations as continuous vectors rather than single discrete classes.

As the first study of *embedding* discourse relations into high dimensional continuous spaces, we mainly focus on social media. Social media is a challenging domain because it contains many acronyms, emojis, unicode, and informal variations of grammatical structure, but its personal nature provides diverse and psychologically-relevant dis-

course patterns which are not often found from newswire text. According to our best knowledge, there are only relatively small datasets for specific types of discourse relations for causal relation (Son et al., 2018) and counterfactual relations (Son et al., 2017), but they are not diverse and large enough to learn general discourse relations.

In this paper, we propose a novel weakly supervised learning method for deriving discourse relation embeddings on social media. We created a social media discourse relation dataset and validated our new approach. Furthermore, we conducted visual investigations on continuous discourse relation spaces and thorough qualitative analysis on the behaviors of DiscRE in both PDTB and social media. Then, we also validated how well our learning method can generalize across different domains by applying DiscRE as transfer learning features for discourse relation downstream tasks.

Our contributions include: (1) a novel model structure which, when weakly supervised, creates embeddings capturing discourse relations (DiscRE), (2) the creation of new Twitter discourse relation dataset and the validation of our approach for the discourse relation classification on the dataset, (3) quantitative and qualitative evaluation of DiscRE on PTDB and downstream social media discourse relation tasks in which DiscRE outperformed strong modern contextual word and sentence embeddings, obtaining a new state-of-the-art performance for causality and counterfactuals, and (4) the release of all of our datasets and models.

2   Related Work

Our work builds on previous studies in discourse relations with two key distinctions: (1) the predominant set of work on discourse relations has focused on annotated newswire datasets (PDTB and RST DT) rather than social media; (2) work to improve discourse parsing has focused either on feature engineering or models for better predicting *predefined* discourse relations rather than embeddings (or latent relations). Such work takes pre-segmented clauses as input (Pitler et al., 2009; Park and Cardie, 2012) or builds full end-to-end discourse parsers (Ji and Eisenstein, 2014; Lin et al., 2014). Kishimoto et al. (2020) looked into adapting BERT for relation classification by pretraining with domain text and connective prediction. Other methods have zeroed-in on implicit discourse relations (those without a connective token) and also used a hierarchical

model but for discourse classification rather than embedding (Bai and Hai, 2018). Some work from Varia et al. (2019); Ma et al. (2021); Zhang et al. (2021) leverage CNNs and graph networks to capture relationships between adjacent discourse units for implicit discourse relation classification.

Some have studied *single* discourse relations over social media. Son et al. (2017) used a hybrid rule-based and feature based supervised classifier to capture counterfactual statements from tweets. Bhatia et al. (2015) and Ji and Smith (2017) applied RST discourse parsing to social media movie review sentiment analysis, showing a pretrained model which was optimized for RST DT, suffered from domain differences when it was run on different domains (e.g., legislative bill). Son et al. (2018) developed a causal relation extraction model using hierarchical RNNs to parse social media. In general, hierarchical RNN-based models have worked well in general for capturing specific relations in social media and other discourse relations outside social media (Bhatia et al., 2015; Son et al., 2018; Ji and Smith, 2017).

Our work is related to modern multi-purpose contextual word embeddings (Devlin et al., 2018; Peters et al., 2018) in the motivation to utilize latent representations in order to capture context-specific meaning. However, our model generates contextual discourse relation embeddings by learning probabilities rather than discrete labels and, it can learn all possible relations even from the same text leveraging posterior probabilities from well-established study (Prasad et al., 2008).

We also build on research that has assembled custom discourse relation datasets or created training instances from existing datasets using discourse connectives (Jernite et al., 2017; Nie et al., 2019; Sileo et al., 2019). Jernite et al. (2017) designed an objective function to learn discourse relation categories (conjunction) based on discourse connectives along with other discourse coherence measurements while Nie et al. (2019) and Sileo et al. (2019) used objectives to predict discourse connectives. Here, we devised an objective function for learning posterior probabilities of discourse relations of the given discourse connectives, so the model can capture more fine-grained senses and discourse relation properties of the connectives[1]. Also, all of them used sentence encoders to learn sentence

---

[1]e.g., 'since' can signal a temporal relation in 'I have been working for this company since I graduated', but might signal a causal relation 'I like him since he is very kind to me'.
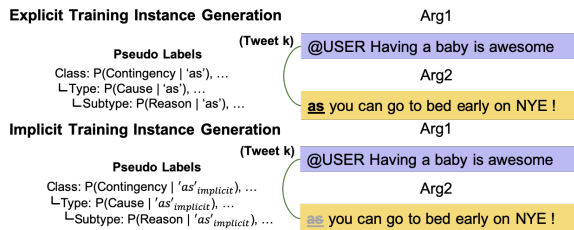
Figure 2: Training instance generation example. For explicit relation training, the training instance is labeled with the posterior probabilities of all possible *Class*, *Type*, and *Subtype* given the explicit connective 'as' from PDTB.
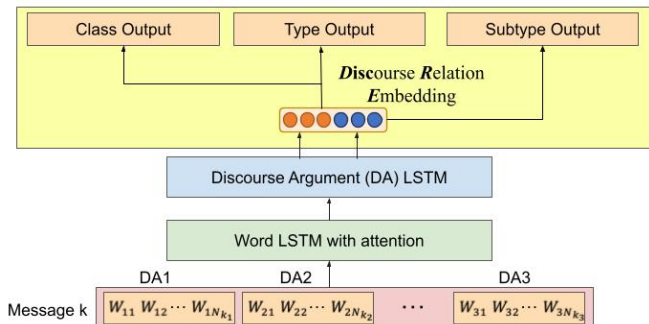

Figure 3: Our model learns different nuances and high dimensional contextual discourse relations by learning probabilities of all possible discourse relations in the relation hierarchy (*Class*, *Type*, and *Subtype*).

representations and compared their learned representations with other state-of-the-art sentence embeddings such as Infersent (Conneau et al., 2017). However, our DiscRE model learns a "discourse relation" representation (i.e. embedding) between discourse arguments rather than the representation of a respective text span of the pair (Figure 1).

Finally, some have studied an RNN-attention-based approach to multitask learning for discourse relation predictions in PDTB (Lan et al., 2017; Ji et al., 2016) and a sentence encoder with multi-purpose learning for discourse-based objectives (Jernite et al., 2017). Also, Liu et al. (2016) leveraged a multi-task neural network for discourse parsing across existing discourse trees and discourse connectives. Shi and Demberg (2019) used next sentence prediction to get better at implicit discourse relation classification.

A particular challenge of these prior works has been to improve performance when no connective is explicitly mentioned in the text. All of these works utilized predefined discrete classes of possible discourse relations. While we were inspired and build on some of their techniques, our task is more broadly defined as producing vector representations of the relationship between discourse segments *not limited to predefined discourse relations* (whether defined with explicit connectives or conventional discourse signals exist or not) and is evaluated over a broad diversity of discourse relation tasks as well as downstream applications.

3 Methods

The base for our model is a hierarchical BiRNN, following work on capturing causal relations in social media (Son et al., 2018), but we have added word-level attention, reflecting the necessity to keep word-level markers while parsing higher-

order discourse relations (e.g., word pairs, modality, or N-grams) (Pitler et al., 2009).

3.1 Data Collection

DiscRE Weakly-Supervised Learning Training Set. No existing annotated discourse relation dataset exists for social media. Thus, we collected random tweets from December 2018 through January 2019 for training. Non-English tweets were filtered out, and URLs and user mentions were replaced with separate special tokens respectively. For training, we collected only messages which contained at least one of the most frequent discourse connectives from each PDTB discourse sense (*Type*) annotation[2] among random tweets from January 2019: up to 3,000 messages for each type of discourse relation which is similar to the numbers in existing social media discourse relation datasets. With this process, we 1) balance our training set to have similar effect sizes of target datasets, 2) minimize potential biases towards a few dominant discourse relations in Twitter, and 3) keep the minimal numbers of discourse relation data samples to validate the effectiveness of the computationally efficient objective function for directly capturing discourse relations. Originally we found 20,787 tweets with our keyword search, but our discourse connective disambiguation process (see details in Section 3.2) left us 11,517 tweets. We chose random 10% of them as our development set to tune hyperparameters.

Qualitative Analysis Evaluation Set. For our qualitative analysis, we separately collected 10,000 random tweets from December 2018 without any

---

[2]after, before, when, but, though, nevertheless, however, because, if, and, for example, or, except, also.

restrictions so we can test our model on an unseen and unbiased natural social media test set as possible. This setting also allows us to conduct qualitative analysis with minimized potential biases which might exaggerate the capabilities of our model (e.g., our model would be evaluated on discourse relations and discourse connectives it had never seen during its training, so it would not be able to depend only on posterior probabilities of certain discourse connectives used as keywords for training set collection to obtain coherent qualitative analysis results).

PDTB-style Twitter Discourse Relation Dataset. As an additional social media evaluation, we created a Twitter discourse relation classification dataset. We collected 360 tweets from September 2020 using the same preprocessing methods for DiscRE training set. Specifically, first, we collected 30 tweets using all discourse connectives of each discourse relation class (i.e., *Contingency*, *Temporal*, *Comparison*, and *Expansion*) as search keywords from random tweets, so 120 tweets in total. Then, three well-trained annotators annotated whether each set of 30 tweets have its target relations as a binary classification. Finally, we randomly shuffled 120 keyword tweets and 240 non-keyword random tweets, and annotators classified four discourse relation classes. Pairwise inter-rater agreement was 85%, with three-way reliable in the moderate range (Fleiss $\kappa = 0.49$). We used majority vote as our discourse relation labels. Among 360 tweets, there were 36 *Contingency*, 8 *Temporal*, 22 *Comparison*, and 43 *Expansion* relations. The rest of the tweets were annotated with *None*.

3.2 Discourse Argument Extraction

We adopted the PDTB-style argument extraction method as it is relatively simple and thus more robust in noisy texts of social media. For argument extraction, we combined approaches of Biran and McKeown (2015) and Son et al. (2018).

We extract the sentences and use the Tweebo parser Kong et al. (2014) to extract discourse arguments (we identified discourse connectives only if there are verb phrases[3]). If there is discourse connective in a sentence, we identify an argument to which a discourse connective attached as *Arg2*, and the other as *Arg1* (Prasad et al., 2007). For discourse connectives at the beginning of a tweet, we identify the text from the beginning until the end

---

[3]minimal discourse units defined in Prasad et al. (2008)

of the first verb phrase separated by punctuation Tweet POS tags or other discourse connectives as *Arg2*, and the rest as *Arg1*; if a discourse connective or coordinating conjunction Tweet POS tag is in the middle, we identify the text from start to the middle connective as *Arg1*, and from the connective to the end as *Arg2* (Biran and McKeown, 2015). We also identify emojis as separate discourse arguments as suggested by (Son et al., 2018) since they play a critical role in signaling implicit relations.

3.3 Training

We use weakly supervised multitask learning with a hierarchy of PDTB-style discourse relation learners (Figure 2). Note that this method, as opposed to entirely self-supervised (i.e. predict next discourse argument), enables us to capture the relationships beyond the likelihood of one discourse argument to appear after another (i.e. how BERT models sentences), which would not necessarily distinguish one relationship from another.

Pseudo Labeling and Training Instance Generation. For each discourse argument pair, the discourse connectives were extracted, and the pair was labelled with all of the possible relations that are found in PDTB. We use the ratio of these possible discourse relations given the discourse connective as a weight within binary cross-entropy loss – this idea of using probabilistic labels follows the work in *pseudo labeling* for image recognition (Lee, 2013). More specifically, two types of training instances were used for the weakly supervised learning of DiscRE: explicit relation pairs and implicit relation pairs. For explicit relation training pairs, the discourse argument which contains discourse connectives is defined as *Arg2* and the rest text span of the pair is defined as *Arg1*. This segmentation method obtained state-of-the-art performances for previous discourse relation tasks (Biran and McKeown, 2015; Son et al., 2018). For implicit relation training pairs, the discourse connective is removed from *Arg2* of each pair; Rutherford and Xue (2015) found this approach can learn strong additional signals quite well, although it is not perfectly equivalent to learning implicit discourse relations[4]. Next, each of these generated pairs were input along with its whole tweet as its context to our DiscRE model

---

[4]Among the discourse connectives we used for our training, only 'if' belongs to the *'Non-omissible'* discourse connective class and even this class showed relatively high effectiveness for implicit relation training when omitted (Rutherford and Xue, 2015).

optimize the model towards the objective function to learn the posterior distributions of all possible relations given the discourse connective in PDTB (Figure 3). Importantly, this mode of labeling is self scalable, yet it also enables a relatively delicate learning objective which considers all possible discourse relations rather than predicting just discourse connectives.

3.4 Discourse Relation Embeddings

We used a hierarchical bidirectional LSTM model; the first layer LSTM (Word LSTM) captures interaction between words of each discourse argument with attention. The second layer LSTM (Discourse Argument LSTM) captures relations among all discourse arguments across the whole tweet. This architecture was inspired by Son et al. (2018) and Ji and Smith (2017) as they found that their similar hierarchical model architecture performed well in related discourse relation tasks. As the first work to attempt embedding relations, we choose RNNs because the sequences of discourse units are of a similar size as where RNNs have been successful over transformers elsewhere (Matero and Schwartz, 2020). Discourse relations, by their definition, describe relations between neighboring or close discourse units, and thus do not have the same motivations for attention-based architectures as long distance dependencies in sequences of words.

This model was optimized on each tweet for training towards the following objective function:

$$J(\theta) = -\sum_i \sum_{j=1}^{N_i} w_{ij} y_{ij} log(f_i(x_{ij})))$$

where i is three levels of discourse relation hierarchy from PDTB (*Class*, *Type*, and *Subtype*) and N_i is the dimension of all existing relations in each level and w_{ij} is the posterior from PDTB of the relations given the discourse connective in the current pair of arguments. This can be viewed as multitask learning of shared RNN layers for three different level outputs (Figure 3). The hidden vectors of *Arg1* and *Arg2* from Discourse Argument LSTM were concatenated to learn *Class* output and *Type* output, as these are relations between two arguments. Whereas, only the hidden vector of *Arg2* from Discourse Argument LSTM was used for learning *Subtype* as it is rather a role of *Arg2*, given the *Class* and *Type* relations (Figure 3). There is a dropout layer with a dropout rate of 0.3 (as suggested in Ji and Smith (2017) and Son et al. (2018))

between Word LSTM and Discourse Argument LSTM.

Finally, for generating DiscRE, the hidden vectors of *Arg1* and *Arg2*, and the output vectors of *Class*, *Type*, and *Subtype* were concatenated. With this structure, DiscRE can capture latent features of discourse relations between any given argument pair, based on the context across all other discourse arguments in addition to probabilities of predefined discourse relations with contextual nuances (Figure 3).

Model Configuration. DiscRE is implemented in PyTorch (Paszke et al., 2019). For hyperparameter tuning, we explored the dimensions of pretrained word embeddings (Glove) and hidden vectors 25, 50, 100, and 200 with SGD and Adam (Kingma and Ba, 2014). We chose the models which obtain best performances on our development set, which used Adam with 200 dimensions and typically 50 epochs. We implemented a word-level attention as defined in (Yang et al., 2016) but with ReLU function for its activation. We compare with other similar models such as: (1) BERT, for which we used BERT base uncased model (12 layers, 768 hidden dimensions, and 12 heads) by HuggingFace [5] and (2) InferSent, for which we used a pretrained model trained with 300 dimension glove vectors as inputs and 2,048 LSTM hidden dimensions.

4 Results

DiscRE was validated on both newswire and social media discourse relation tasks. Additionally, qualitative analysis on the DiscRE representations were explore for both the domains.

4.1 Evaluations

First, we examined whether DiscRE can capture discourse relations in PDTB, even though grammatical properties and general text formats of newswire and social media are quite different. Then, we evaluated our model for social media discourse relation tasks: causal relation prediction and Twitter discourse relation classification. We used linear SVMs for all transfer learner classifiers for evaluation as this model obtained the best performance from the previous related work (Son et al., 2018).

---

[5]https://huggingface.co/
bert-base-uncased

Models	CON.	TEM.	COM.	EXP.	Mic.	Mac.
Ngrams	0.575	0.693	0.757	0.757	0.709	0.695
BERT	**0.612**	0.724	0.746	0.748	0.714	0.708
Inferse.	0.604	0.670	0.738	0.726	0.693	0.685
DiscRE	0.598	**0.736**	**0.768**	**0.768**	**0.726**	**0.718**

Table 1: F1 scores of the four-way PDTB discourse class prediction ('CON.': *Contingency*, 'TEM.': *Temporal*, 'COM.': *Comparison*, 'EXP.': *Expansion*). We report both micro F1 and macro F1. DiscRE obtained the best performances across all four discourse relation classes except for the second best performance for Contingency class prediction F1.

Transfer Learning on PDTB. In order to measure how well our model can generalize to different domains and capture predefined newswire discourse relations, we conducted transfer learning experiments for predicting the four senses of Level-1 discourse relation classes: *Contingency*, *Temporal*, *Comparison*, and *Expansion*.

We extract DiscRE from the pairs: *Arg1* and *Arg2* from the PDTB dataset, and used them as transfer learning features to a linear classifier. The PDTB dataset was created with the annotators first segmenting the texts into discourse arguments, and then annotating a discourse relation between each pair of neighboring discourse arguments (marked as *Arg1* and *Arg2*). To make a fair comparison, we extracted BERT, Ngrams, and Infersent from *Arg1* and *Arg2* and the concatenation of *Arg1* and *Arg2* to use as separate features, so that the transfer learned model can recognize the notion of *Arg1* and *Arg2* and utilize the whole context as well. The classifiers were trained with each of these embeddings and we report the performances.

As suggested in Prasad et al. (2007), we used Sections 2 to 21 for training and Section 23 for testing in PDTB. Despite the relatively small number of the training set and larger domain differences with newswire target domains in its pretraining procedures, DiscRE still obtained the best performance for overall discourse relation predictions except for *Contingency* classification F1. This may indicate that DiscRE learns domain-agnostic signals for discourse relations leveraging discourse connectives in the weakly supervised multitask learning settings. (Table 1).

Causal Relation Prediction on Social Media. We evaluated our model on a causality prediction task on social media messages collected by Son et al. (2018). The DiscRE embeddings of the messages were extracted and for each message, the embeddings were averaged over for the transfer

Model	F1
(Son et al., 2018)	0.791
BERT	0.746
Infersent	0.709
DiscRE	0.752
BERT Fine-Tuned	0.789
DiscRE + ALL	**0.807**

Table 2: Causality prediction performance of DiscRE compared to other models. DiscRE-based classifier obtained the new state-of-the-art performance.

learning features for causality prediction. For comparison, BERT embeddings were also extracted for each discourse unit, and averaged for each message in the dataset, and Infersent sentence embeddings were directly extracted from the messages. The transfer learned classifier from DiscRE embeddings can be used to improve over the best results reported in the previous work on causality prediction (Son et al., 2018). DiscRE obtained better performances ($F1 = 0.752$) than BERT ($F1 = 0.746$) and Infersent (F1=0.709) and overall, this simple transfer learning approach using obtained a comparable performance to the models used in Son et al. (2018) ($F1 = 0.791$) (Table 2). On further exploration, we found that fine-tuning BERT for the causality prediction task improved the performance to $F1 = 0.789$. Furthermore, when DiscRE was used along with best performing text features from Son et al. (2018) (N-grams, Tweet POS tags, Word Pairs (Pitler et al., 2009), sentiment tags) of the messages for transfer learning, we obtained a new state-of-the-art performance (See Table 2)

---

[6] Interestingly, on Twitter, the attention weights of social-media-specific variations of 'because' obtained similar weights even though the DiscRE model was not systematically designed to capture domain differences of discourse connectives: 'because': 0.16, 'bcuz': 0.18, 'cos': 0.16, 'cuz': 0.15, 'cause': 0.16.

Models	CON.	TEM.	COM.	EXP.	None	Mic.	Mac.
Ngrams	0.386	0.386	0.353	0.119	0.813	0.686	0.407
BERT	0.412	0.000	0.426	0.086	0.857	0.706	0.316
Inferse.	0.390	0.111	0.566	0.324	0.867	0.719	0.452
DiscRE	**0.478**	**0.421**	**0.591**	**0.400**	**0.883**	**0.758**	**0.554**

Table 3: F1 scores of the discourse class prediction on Twitter ('CON.': *Contingency*, 'TEM.': *Temporal*, 'COM.': *Comparison*, 'EXP.': *Expansion*). Then, we report both micro F1 and macro F1. DiscRE obtained the best performance across all relations.

(a) Average Attention Weights on Twitter
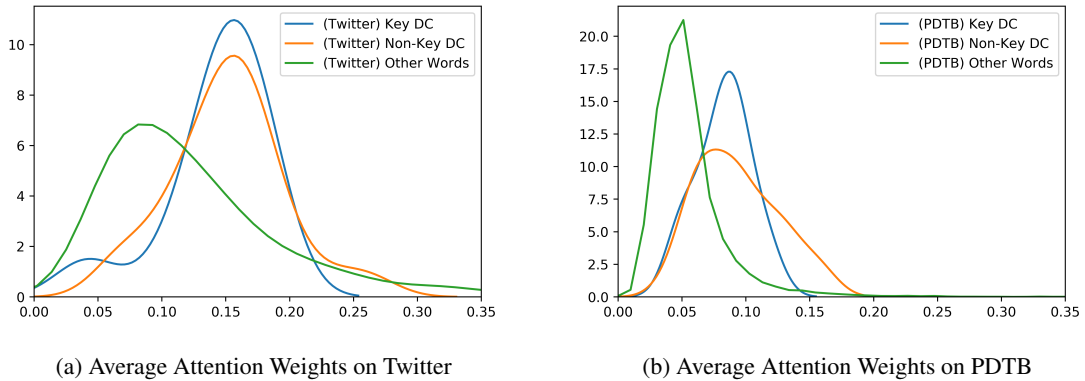
(b) Average Attention Weights on PDTB

Figure 4: Distribution plot with attention weights as a variable in x-axis, 'Key DC': discourse connectives used as keywords for the training set collection, 'Non-Key DC': discourse connectives which were not included in the keywords. We analyzed the average attention weight distributions of discourse connectives vs other words. Discourse connectives tend to receive higher attention on both PDTB and Twitter[6].

Discourse Relation Classification on Social Media. To validate DiscRE beyond the existing corpus of newswire domain, it was applied to a discourse relation classification task on our new Twitter discourse relation dataset. We extracted DiscRE, BERT, Ngrams, and Infersent from tweets with the same methods used in the causality task. We conducted 10-fold cross validation and report F1 scores of the models on each class in Table 3. The result showed that DiscRE obtains the best performance across all the classes. (Micro F1=0.758).

4.2 Qualitative Analysis on DiscRE model

Attention Analysis. First, we ran pretrained DiscRE on the evaluation tweet dataset (Section 3.1) and investigated average attention weights. Discourse connectives gained higher attention than non-discourse-connective words (Figure 4).[7] This suggests that discourse connectives play a quite significant role in DiscRE.

Furthermore, we observed that both, the dis-

course connectives used as keywords for training set collection, as well as the relatively less frequent discourse connectives obtained higher attention weights than other words on the random tweet evaluation set. This pattern supports that our model was not biased towards only prevailing discourse connectives it has already seen from the training set, but generalized quite well on unseen discourse connectives.

When we analyzed attention weights on the DiscRE model for the PDTB dataset, it showed a similar pattern. Although all words in the PDTB vocabulary generally obtained lower attention, the discourse connectives still obtained higher attention weights than other words, and relatively high attention weights were distributed on both keyword and non-keyword discourse connectives in PDTB as well. These results suggest that DiscRE can capture words with important discourse signals even on the other domains.

DiscRE Analysis. We evaluated DiscRE on social media discourse relations datasets which are publicly available: causality (Son et al., 2017) and

---

[7]Beyond some outliers due to noisy unigrams and social-media-specific discourse arguments (e.g., emojis or verb phrases with omitted subjects)

counterfactual (Son et al., 2018). We averaged the DiscRE embeddings of all adjacent pairs of discourse arguments per message and visualized using tSNE (Figures 5, 6). In general, discourse relations are diverse and even the same *Type* show up in various different forms in both explicit and implicit relations, so the distinctions between them are very hard to be captured within just two dimensions. Nevertheless, we found fairly clear patterns that distinguish two different discourse relations; majority counterfactual messages tend to cluster separately towards the left, as compared to causality messages (Figure 5). *Conjunctive Normal* and *Conjunctive Converse* forms of counterfactuals are especially clustered at the left side separately (e.g., "I would be healthier, if I had worked out regularly") (Son et al., 2017).

It is noteworthy that the counterfactual relation does not exist as a discourse relation tag in PDTB, but DiscRE still captures its distinguishable properties and even different forms of it (i.e., *Wish verb* forms and *Conjunctive* forms). While this visualization provides significant insights about semantic differences of discourse relations, further analysis over coherent clusters helps us see some discourse-based properties in common (e.g., see 'Message A' and 'Message B' on Figure 5).

Additionally, we investigated how well DiscRE can generalize to newswire domain by projecting DiscRE embeddings of discourse relations in the PDTB testset into 2D tSNE, similar to the visualization of causal and counterfatual relations (Figure 6). Even though we used most coarse-grained discourse relation classes, DiscRE captured quite coherent patterns of clusters for different relations. Nevertheless, many implicit discourse relations were clustered together on the upper left part as they are generally harder to be captured (Pitler et al., 2008; Rutherford and Xue, 2015).

5 Conclusion

This paper suggests a difference in how semantics is modeled in NLP, moving beyond word-level embeddings to embeddings that capture the semantics of discourse relations. We explored a new task of creating latent discourse relation embeddings, designing a novel weakly supervised multitask learning method and evaluating it both quantitatively and qualitatively over social media and newswire domains. While we built on previous work over discourse relation classes, our results suggest the *con-*
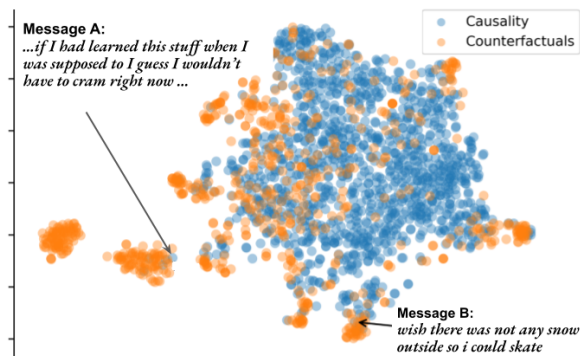

Figure 5: DiscRE differences between counterfactual messages and causality messages. Counterfactual messages are generally positioned at the left side compared to causality messages. When we investigated edge cases of causality messages that clustered closely with counterfactuals, we found causality messages which contained counterfactual relations inside ('Message A': 'is doing great.... lol. If I had learned this stuff when I was supposed to I guess I wouldn't have to cram right now. Oh well. There's always next year... or grade 12.' 'Message B': 'i wish there was not any snow outside so i could skate').

tinuous discourse relation embeddings (DiscRE) has certain benefits over manual categorizations. Continuous representations of relations between segments of text have been relatively unexplored yet they can yield subtle attributes of discourse relations, yielding strong performance in applications and perhaps new organizations of functional discourse relations.

Our model obtained the best performance on the discourse relation classification tasks in both PDTB and our new Twitter discourse dataset. Our model also obtained a new state-of-the-art performance using DiscRE in the social media causal relation prediction task. Further, for predicting discourse relations over PDTB, we found DiscRE achieved the higher performance than other embeddings, suggesting a focus on embedding *relations* can capture information not available in other types of modern embeddings which focus on representing particular word or phrase instances rather than their relationships. We release our dataset, code and pretrained models, for others to explore this new task, better develop continuous representations of discourse relations, as well as to extend discourse relation parsing beyond newswire to other domains.
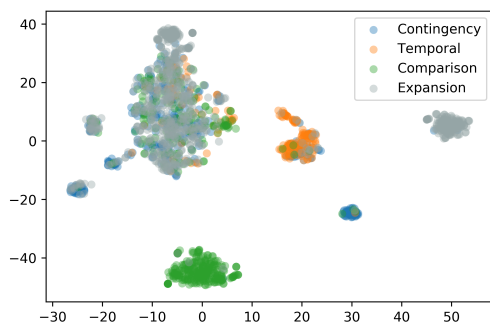
Figure 6: DiscRE differences between the four discourse relation classes of the PDTB dataset. Many examples of implicit discourse relations were clustered on the upper left side. Expansion is a quite general class which may overlap semantically with other types of relations, so they were more widely spread than other relations.

6 Limitations

The model delineated in this work is scalable with large amounts of unsupervised data, but still orders of magnitude less than what modern language models require. The social media validation was performed on a small annotated dataset with a high inter-annotator agreement, limited to 360 tweets that had examples from each relation class . The model was trained on a single 12GB memory GPU (we used a NVIDIA Titan XP graphics card). The approach should be expected to work best with languages that have limited morphology, like English.

The weakly supervised approach has a small limitation in that it still aligns the model, to some degree, with an existing tagset (i.e. the PDTB discourse relation tagset), but our results suggested we were able to capture relations beyond it (e.g. capturing a relation that is a mix of causal explanation and counterfactuals).

7 Ethical Considerations

All of our work is restricted to document-level information; No user-level information is used.

References

Hongxiao Bai and Zhao Hai. 2018. Deep enhanced representation for implicit discourse relation recognition. In *COLING*.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218.

Or Biran and Kathleen McKeown. 2015. Pdtb discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 96–104.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yacine Jernite, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

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

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. In *Proceedings of NAACL-HLT*, pages 332–342.

Yangfeng Ji and Noah A Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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 12th Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar. Association for Computational Linguistics.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308.

Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Yuhao Ma, Yu Yan, and Jie Liu. 2021. *Implicit Discourse Relation Classification Based on Semantic Graph Attention Networks*. Association for Computing Machinery, New York, NY, USA.

Matthew Matero and H. Andrew Schwartz. 2020. Autoregressive affective language forecasting: A self-supervised task. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2913–2923, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 683–691. Association for Computational Linguistics.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884.

Diana Nicoleta Popa, Julien Perez, James Henderson, and Eric Gaussier. 2019. Implicit discourse relation classification with syntax-aware contextualized word representations. In *The Thirty-Second International Flairs Conference*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5790–5796, Hong Kong, China. Association for Computational Linguistics.

Damien Sileo, Tim Van de Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. 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 3477–3486.

Youngseo Son, Nipun Bayas, and H Andrew Schwartz. 2018. Causal explanation analysis on social media. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3350–3359.

Youngseo Son, Anneke Buffone, Joe Raso, Allegra Larche, Anthony Janocko, Kevin Zembroski, H Andrew Schwartz, and Lyle Ungar. 2017. Recognizing counterfactual thinking in social media texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 654–658.

Siddharth Varia, Christopher Hidey, and Tuhin Chakrabarty. 2019. Discourse relation prediction: Revisiting word pairs with convolutional networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 442–452, Stockholm, Sweden. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Yingxue Zhang, Fandong Meng, Peng Li, Ping Jian, and Jie Zhou. 2021. Context tracking network: Graph-based context modeling for implicit discourse relation recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1592–1599, Online. Association for Computational Linguistics.