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

Youngseo Son and H. Andrew Schwartz
Stony Brook University
Stony Brook, NY
{yson,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 limits the universe of potential relationships and their nuanced differences. Analogous to 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 *between* two segments, often with no word or phrase in that gap) which presents 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 to learn diverse and nuanced relations between discourse segments in social media. Results show DiscRE can: (1) obtain the best performance on Twitter discourse relation classification task (macro F1=0.76) (2) improve the state of the art in social media causality prediction (from $F1 = .79$ to .81), (3) perform beyond modern sentence and contextual word embeddings at traditional discourse relation classification, and (4) capture novel nuanced relations (e.g. relations semantically 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

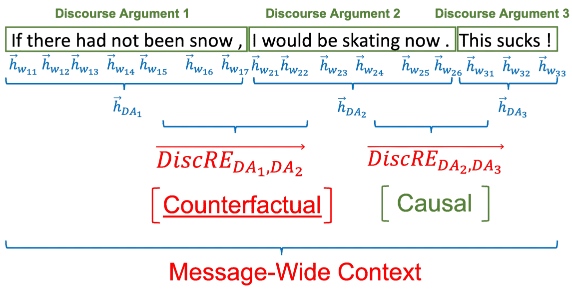


Figure 1: Our model DiscRE predicts relations of adjacent discourse arguments based on other text spans of the whole message as context. By learning and embedding fine-grained properties of discourse relation with the posteriors from PDTB into a continuous vector space, DiscRE may learn existing discourse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

sophisticated schema, annotation is non-trivial prohibiting extensive development in new domains (e.g., social media). Thus, progress in developing, training and evaluating discourse relations approaches has happened over *discrete* models with *labeled newswire* corpus (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 labels but still 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 discourse 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.

Thus, 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) the proposal of a novel model structure which can produce latent discourse relation embeddings, (2) the creation of new Twitter discourse relation dataset and the validation of our approach for the discourse relation classification on the dataset, (3) qualitative analysis on model predictions, (4) evaluations on downstream social media discourse relation tasks in which DiscRE outperformed strong modern contextual word and sentence embeddings and obtained a new state-of-the-art performance, (5) validation analysis on PDTB, and (6) the release of all of our datasets and models.

2 Related Work

Most researchers trained and evaluated their discourse relation parsers on the annotated newswire dataset (PDTB and RST DT); some researchers focused on feature engineering for discourse relation predictions with an assumption that they were given segmented discourse argument (Pitler et al., 2009; Park and Cardie, 2012) while others built full end-to-end discourse parsers (Ji and Eisenstein, 2014; Lin et al., 2014). Bhatia et al. (2015) and Ji and Smith (2017) applied RST discourse parsing to social media movie review sentiment analysis, but discourse structure was built by pretrained model which is optimized for RST DT, so the model suffered from domain differences when it was run on documents which have drastically different formats (e.g., legislative bill).

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. Son et al. (2018) developed a causal relation extraction model using hierarchical RNNs to parse social media. Since hierarchical RNN-based models have worked well in general for capturing specific relations in social media and other discourse relations outside social media (Son et al., 2018; Ji and Smith, 2017; Bhatia et al., 2015), we developed our weakly supervised, multi-class discourse relation embeddings around a hierarchical bidirectional RNN model with word-level attention for discourse relation parsing on Twitter. Since tweets tend to have many noisy features (e.g., hashtags, URL, or dropping subjects) and informal grammar in short length, it is difficult to obtain accurate RST-style comprehensive hierarchical discourse structures. Thus, we also employed PDTB-style method in which discourse relation is predicted between only adjacent discourse arguments, but we capture the context across all other discourse arguments by using hidden vectors of argument pairs from Discourse Argument LSTM which ran on the whole tweet (Figure 1).

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).

Other researchers collected their own 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

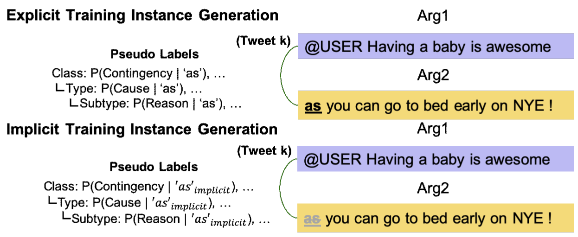


Figure 2: Training instance generation example. For explicit relation training, the training instance is labeled with the posterior probabilities of all possible *Class*, *Type*, and *Subtype* given the explicit connective 'as' from PDTB.

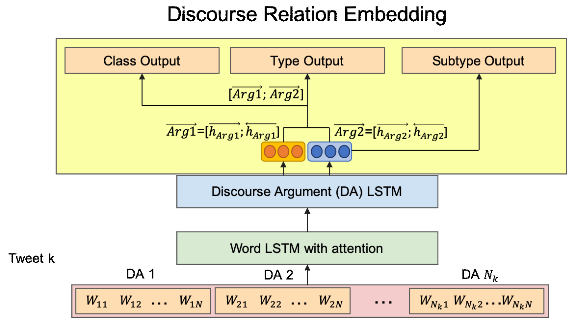


Figure 3: Our model learns different nuances and high dimensional contextual discourse relations by learning probabilities of all possible discourse relations in the relation hierarchy (*Class*, *Type*, and *Subtype*).

of the connectives[1]. Also, all of them used sentence encoders to learn sentence 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 "discourse relation" representation between discourse arguments rather than the representation of a respective text span of the pair (Figure 1).

Finally, some researchers recently studied an RNN-based with attention mechanism with 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. From these sets of prior work, a particular challenge on their main focus 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 either of explicit connectives or conventional discourse signals exist or not).

3 Methods

The basis 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, because many word-level features have been useful for discourse relation parsing (e.g., word pairs, modality, or N-grams) (Pitler et al., 2009).

3.1 Data Collection

DiscRE Weakly-Supervised Learning Training Set. There was no existing annotated discourse relation dataset in social media. Therefore, we collected random tweets from December 2018 through January 2019 for training. For preprocessing, we filtered out non-English tweets. Also, we replaced URLs and user mentions 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.

[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'.

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

Qualitative Analysis Evaluation Set. For our qualitative analysis, we separately collected 10,000 random tweets from December 2018 without any 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 [3]. We also added 240 random tweets without using any keywords. 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 *Contigency*, 8 *Temporal*, 22 *Comparison*, and 43 *Expansion* relations.

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 all sentences and if there is discourse connective inside 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 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). Also, we identify emojis as separate discourse arguments as suggested by (Son et al., 2018) since they plays a critical role for signaling implicit relations.

For discourse connective disambiguation[4], we identified discourse connectives only if there are verb phrases[5] before and after them as Son et al. (2018) found that this simplified method was effective for capturing social media discourse 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. We extract discourse connectives of a given discourse argument pair and label that pair with all of the possible relations that are found in PDTB. Then, we use the ratio of those possible discourse relations given the discourse connective as a weight within binary cross-entropy loss – this idea of using probabilistic labels follow work in *pseudo labeling* for image recognition (Lee, 2013). More specifically, we generated two types of training instances for the weekly supervised learning of DiscRE: explicit relation pairs and implicit relation pairs. For explicit relation training pairs, we define the discourse argument which contains discourse connectives as *Arg2* and the rest text span of the pair as *Arg1* as this segmentation method obtained state-of-the-art performances for previous discourse relation tasks (Biran and McKe-

[3] e.g., whether 30 tweets collected using discourse connnectives of contigency actually contain contigency relations?

[4] e.g., 'and' can be used as a discourse connective ('I fell asleep again **and** I got late') or a simple connections of two words ('I want apple **and** banana').

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

own, 2015; Son et al., 2018). For implicit relation training pairs, we remove the discourse connective 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[6]. Then, we input each of these generated pairs along with its whole tweet as its context to our DiscRE model 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 and 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.

Then, we optimized this model 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). We concatenate the hidden vectors of *Arg1* and *Arg2* from Discourse Argument LSTM to learn *Class* output and *Type* output as these are relations between two arguments while we use only the hidden vector of *Arg2* from Discourse Argument LSTM for learning *Subtype* as it is rather a role of *Arg2* given the *Class* and *Type* relations (Figure 3). We put dropout layer between Word LSTM and Discourse Argument LSTM and used 0.3 for its rate as suggested in Ji and Smith (2017) and Son et al. (2018).

Finally, for generating DiscRE, we concatenated the hidden vectors of *Arg1* and *Arg2* to the concatenation of output vectors of *Class*, *Type*, and *Subtype*. With this structure, DiscRE can capture a latent features of discourse relations between the 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. Our DiscRE model is implemented in PyTorch (Paszke et al., 2019). For hyperparameter tuning, we explored both dimensions of pretrained word embeddings[7] 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 within at most 1,000 epochs[8]. We implemented a word-level attention as defined in (Yang et al., 2016) but with ReLU function for its activation. For BERT extraction, we used BERT base uncased model (12 layers, 768 hidden dimensions, and 12 heads) by HuggingFace and for infersent we used a pretrained model trained with 300 dimension glove vectors as inputs and 2,048 LSTM hidden dimensions[9].

4 Results

We validated DiscRE on both newswire and social media discourse relation tasks. Then, we conducted qualitative analysis on the internal representations of our model and its DiscRE prediction on both Twitter and PDTB.

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

[6]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).

[7]http://nlp.stanford.edu/data/glove.twitter.27B.zip

[8]200 dimension with Adam was the best setting for DiscRE.

[9]https://github.com/facebookresearch/InferSent

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-wary PDTB discourse class prediction ('CON.': *Contingency*, 'TEM.': *Temporal*, 'COM.': *Comparison*, 'EXP.': *Expansion*). Then, we report both micro F1 and macro F1. DiscRE obtained the best performances across all four discourse relation clasees except for the second best performance for Contingency class prediction F1.

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.

SVMs for all transfer learner classifiers for evaluation as this model obtained the best performance from the previous related work (Son et al., 2018).

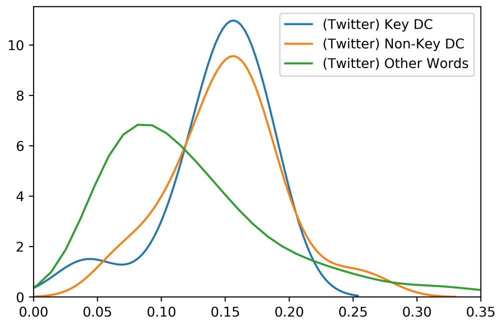
Transfer Learning on PDTB. In order to measure how well our model can generalize to different domains and capture predefined newswire discourse relations, we conducted similar transfer learning experiments for predicting four senses of Level 1 discourse relation classes (*Contingency*,*Temporal*,*Comparison*,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

In PDTB, annotators first segmented texts into discourse arguments, then annotated a discourse relation between each pair of neighboring discourse arguments (marked as *Arg1* and *Arg2*). Therefore, we extracted BERT, Ngrams, and Infersent from *Arg1* and *Arg2* and the concatenation of *Arg1* and *Arg2* and use them as separate features, so the transfer learner model can recognize the notion of *Arg1* and *Arg2* and utilize the whole text context as well. Then, we extract DiscREs from the pairs of *Arg1* and *Arg2* and used them as transfer learning features. Then, we trained classifiers with each of those embeddings and compared the performances.

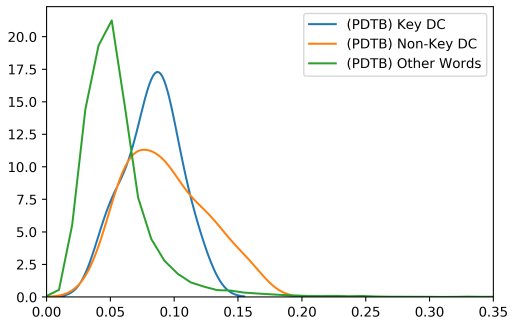
As suggested in Prasad et al. (2007), we used Sections 2 to 21 for training and Section 23 for testing in PDTB. Despite the relatively small number of the training set and larger domain differences with newswire target domains in its pretraining procedures, DiscRE still obtained the best performances 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). We extracted DiscRE and BERT of the messages and average all embeddings in each message as transfer learning features for causality prediction of each message. Then, we extracted Infersent sentence embeddings from messages. Then, we trained classifiers with each of those embeddings and compared the performances as this model obtained the best result from the previous work (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). For further training, we directly fine-tuned BERT for the causality prediction task and the performance increased to F1=0.789. Also, when we used DiscRE 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 the new state-of-the-art performance.

Discourse Relation Classification on Social Media. In order to validate a main objective of DiscRE for extending effective discourse relation parsing beyond the existing corpus of newswire domain, we applied DiscRE 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 previous causality prediction task. We conducted 10 fold cross validations and reported F1 scores of the embedding models on each class. The result showed that DiscRE obtains the best performances across all classes and average F1 scores (Micro F1=0.758).



(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[10].

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 performances across all relations.

4.2 Qualitative Analysis on DiscRE model

Attention Analysis. First, we ran our trained DiscRE model on evaluation tweet dataset and investigated average attention weights of words when they were embedded into DiscRE. Even though there are some outliers due to noisy unigram social-media-specific discourse arguments (e.g., emojis or verb phrases with omitted subjects), generally discourse connectives gained higher attention than non-discourse-connective words (Figure 4). This suggests that discourse connectives play a quite significant role when our model produces DiscRE.

Furthermore, we observed that both discourse connectives used as keywords for training set collection and the other 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 seen from the training set but generalized quite well on unseen discourse relations.

Additionally, when we analyzed attention weights our DiscRE model when it embedded discourse relations on PDTB, it showed the similar pattern. Although all words in the PDTB vocabulary generally obtained lower attention, still discourse connectives obtained higher attention weights than other words and DiscRE distributed relatively high attention weights 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 explored DiscREs on discourse relations in social media which are publicly available: causality (Son et al., 2017) and counterfactual (Son et al., 2018) social media dataset. We averaged all DiscREs of all adjacent pairs of discourse arguments per message and visualized two dimensional tSNE of them. In general, types of discourse relations are diverse and even 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 which distinguish two different discourse relations; majority counterfactual messages tend to cluster separately to the left side compared to causality messages (Figure 5). Especially, *Conjunctive Normal* and *Conjunctive Converse* forms

[10]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.

of counterfactuals are 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, still DiscRE captured distinguishable unique properties of it compared to causal relations and even different forms of counterfactuals (i.e., *Wish verb* forms and *Conjunctive* forms). This visualization analysis provided significant insights about semantic differences of discourse relations, but at the same time, further analysis over coherent clusters helped 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 DiscREs of discourse relations in the PDTB testset into 2D tSNE in the same setting used for 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).

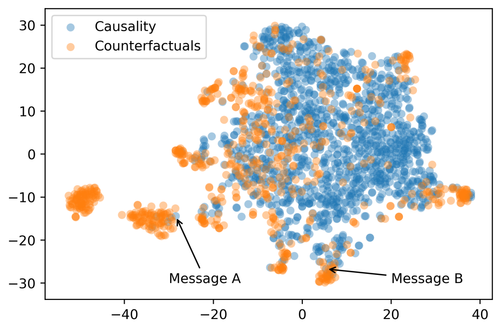


Figure 5: DiscRE differences between counterfactual messages and causality messages. Counterfactual messages are generally positioned at the left side compared to causality messages. When we investigated edge cases of causality messages clustered closely with counterfactuals, we found causality messages which contained counterfactual relations inside ('Message A' and 'Message B'[11].)

[11] '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'.

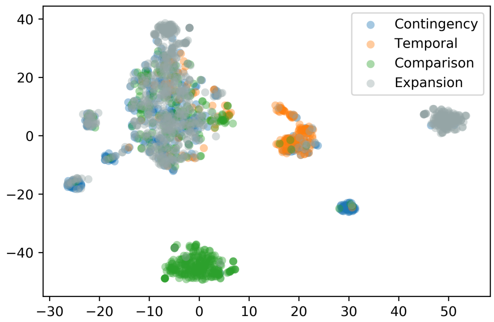


Figure 6: DiscRE differences between Four discourse relation classes of PDTB test set. Many examples of implicit discourse relations were clustered on the upper left side. Expansion is a quite general class which may overlap semantically with other types of relations, so they were more widely spread than other relations.

5 Conclusion

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 *continuous* 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. Also, our model 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* (i.e. the space between text segments) 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.

References

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*.

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*.

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.

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.

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.