Improving Implicit Discourse Relation Recognition with Discourse-specific Word Embeddings

Changxing Wu[1,2], **Xiaodong Shi**[1,2]*, **Yidong Chen**[1,2], **Jinsong Su**[3], **Boli Wang**[1,2]
Fujian Key Lab of the Brain-like Intelligent Systems, Xiamen University, China[1]
School of Information Science and Technology, Xiamen University, China[2]
Xiamen University, China[3]
wcxnlp@163.com boliwang@stu.xmu.edu.cn
{mandel, ydchen, jssu}@xmu.edu.cn

Abstract

We introduce a simple and effective method to learn discourse-specific word embeddings (*DSWE*) for implicit discourse relation recognition. Specifically, *DSWE* is learned by performing connective classification on massive explicit discourse data, and capable of capturing discourse relationships between words. On the PDTB data set, using *DSWE* as features achieves significant improvements over baselines.

1 Introduction

Recognizing discourse relations (e.g., *Contrast, Conjunction*) between two sentences is a crucial subtask of discourse structure analysis. These relations can benefit many downstream NLP tasks, including question answering, machine translation and so on. A discourse relation instance is usually defined as a discourse connective (e.g., *but, and*) taking two arguments (e.g., *clause, sentence*). For explicit discourse relation recognition, using only connectives as features achieves more than 93% in accuracy (Pitler and Nenkova, 2009). Without obvious clues like connectives, implicit discourse relation recognition is still challenging.

The earlier researches usually develop linguistically informed features and use supervised learning method to perform the task (Pitler et al., 2009; Lin et al., 2009; Louis et al., 2010; Rutherford and Xue, 2014; Braud and Denis, 2015). Among these features, word pairs occurring in argument pairs are considered as important features, since they can partially catch discourse relationships between two arguments. For example, synonym word pairs like *(good, great)* may indicate a *Conjunction* relation, while antonym word pairs like *(good, bad)* may mean a *Contrast* relation. However, classifiers based on word pairs in previous work do not work well because of the data sparsity problem. To address this problem, recent researches use word embeddings (aka distributed representations) instead of words as input features, and design various neural networks to capture discourse relationships between arguments (Zhang et al., 2015; Ji and Eisenstein, 2015; Qin et al., 2016; Chen et al., 2016; Liu and Li, 2016). While these researches achieve promising results, they are all based on pre-trained word embeddings ignoring discourse information (e.g., *good, great,* and *bad* are often mapped into close vectors). Intuitively, using word embeddings sensitive to discourse relations would further boost the performance.

In this paper, we propose to learn discourse-specific word embeddings (*DSWE*) from explicit data for implicit discourse relation recognition. Our method is inspired by the observation that synonym (antonym) word pairs tend to appear around the discourse connective *and (but)*. Other connectives can also provide some discourse clues. We expect to encode these discourse clues into the distributed representations of words, to capture discourse relationships between them. To this end, we use a simple neural network to perform connective classification on massive explicit data. Explicit data can be considered to be automatically labeled by connectives. While they cannot be directly used as training data for implicit discourse relation recognition and contain some noise, they are effective enough to provide weakly supervised signals for training the discourse-specific word embeddings.

We apply *DSWE* as features in a supervised neural network for implicit discourse relations recognition. On the PDTB (Prasad et al., 2008), using *DSWE* yields significantly better performance than using off-the-shelf word embeddings,

[*] Corresponding author.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 269–274
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2042

or recent systems incorporating explicit data. We detail our method in Section 2 and evaluate it in Section 3. Conclusions are given in Section 4. Our learned *DSWE* is publicly available at here.

2 Discourse-specific Word Embeddings

In this section, we first introduce the neural network model for learning discourse-specific word embeddings (*DSWE*), and then the way of collecting explicit discourse data for training. Finally, we highlight the differences between our work and the related researches.

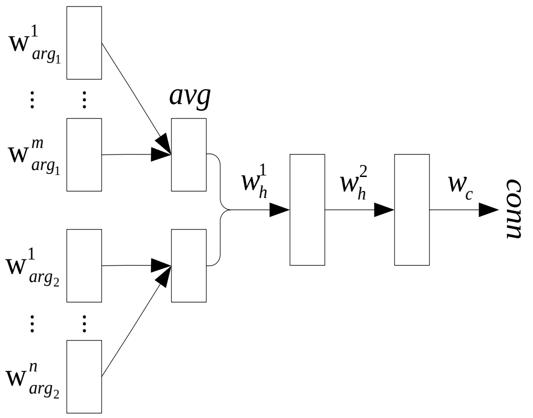

Figure 1: Neural network model for learning *DSWE*. An explicit instance is denoted as $(arg_1, arg_2, conn)$. w$^1_{arg_1}$, ..., w$^m_{arg_1}$ mean the words in arg_1. Two arguments are concatenated as input and the number of hidden layers is not limited to two.

We induce *DSWE* based on explicit data by performing connective classification. The connective classification task predicts which discourse connective is suitable for combining two given arguments. It is essentially similar to implicit relation recognition, just with different output labels. Therefore, any existing neural network model for implicit relation recognition can be easily used for connective classification. We adapt the model in (Wu et al., 2016) for connective classification because it is simple enough to enable us to train on massive data. As illustrated in Figure 1, an argument is first represented as the average of distributed representations of words in it. On the concatenation of two arguments, multiple non-linear hidden layers are then used to capture the interactions between them. Finally, a *softmax* layer is stacked for classification. We combine the cross-entropy error and regularization error multiplied by the coefficient λ as the objective function. During training, we initialize distributed representations of all words randomly and tune them to minimize the objective function. The finally obtained distributed representations of all words are our discourse-specific word embeddings.

Collecting explicit discourse data includes two steps: 1) distinguish whether a connective occurring reflects a discourse relation. For example, the connective *and* can either function as a discourse connective to join two *Conjunction* arguments, or be just used to link two nouns in a phrase. 2) identify the positions of two arguments. According to (Prasad et al., 2008), arg_2 is defined as the argument following a connective, however, arg_1 can be located within the same sentence as the connective, in some previous or following sentence. Lin et al. (2014) show that the accuracy of distinguishing connectives is more than 97%, while identifying arguments is below than 80%. Therefore, we use the existing toolkit[1] to find discourse connectives, and just collect explicit instances using patterns like $[arg1\ because\ arg2]$, where two arguments are in the same sentence, to decrease noise. We believe these simple patterns are enough when using a very large corpus. Note that there are 100 discourse connectives in the PDTB, we ignore four parallel connectives (e.g., *if...then*) for simplicity. The way of collecting explicit data can be easily generalized to other languages, one just need to train a classifier to find discourse connectives following (Lin et al., 2014).

Some aspects of this work are similar to (Biran and McKeown, 2013; Braud and Denis, 2016). Based on massive explicit instances, they first build a word-connective co-occurrence frequency matrix[2], and then weight these raw frequencies. In this way, they represent words in the space of connectives to directly encode their discourse function. The major limitation of their approach is that the dimension of the word representations must be less than or equal to the number of connectives. By comparison, we learn *DSWE* by predicting connectives conditioning on arguments, which yields better performance and has no such dimension limitation. Some researchers use explicit data as additional training data via multi-task learning (Lan et al., 2013; Liu et al., 2016) or data selection (Rutherford and Xue, 2015; Wu et al., 2016).

[1] https://github.com/linziheng/pdtb-parser.

[2] Biran and McKeown (2013) calculate co-occurrences between word pairs and connectives.

In both cases, explicit data are directly used to estimate the parameters of implicit relation classifiers. As a result, it is hard for them to incorporate massive explicit data because of the noise problem. By contrast, we leverage massive explicit data by learning word embeddings from them.

3 Experiments

3.1 Data and Settings

We collect explicit data from the *Xin* and *Ltw* parts of the English Gigaword Corpus (3rd edition), and get about 4.92M explicit instances. We randomly sample 20,000 instances as the development set and the others as the training set for *DSWE*. After discarding words occurring less than 5 times, the size of the vocabulary is 185,048. For the connective classification task, we obtain an accuracy of about 53% on the development set.

We adapt the neural network model described in Figure 1 as the classifier for implicit discourse relation recognition (*CDRR*). Specifically, we concatenate some surface features with the last hidden layer as the input of the *softmax* layer to predict discourse relations. We choose 500 *Production rule* (Lin et al., 2009) and 500 *Brown Cluster Pair* (Rutherford and Xue, 2014) features based on mutual information using the toolkit provided by Peng et al. (2005). Our learned *DSWE* is used as the pre-trained word embeddings for *CDRR*, and fixed during training.

Hyper-parameters for training *DSWE* and *CDRR* are selected based on their corresponding development set, and listed in Table 1.

Hyper-parameter	*DSWE*	*CDRR*
$wdim$	300	300
$hsizes$	[200]	[200, 50]
lr	1.0	0.005
λ	0.0001	0.0001
$update$	SGD	AdaGrad
f	ReLU	ReLU

Table 1: Hyper-parameters for training *DSWE* and *CDRR*. $wdim$ means the dimension of word embeddings, $hsizes$ the sizes of hidden layers, lr the learning rate, λ the regularization coefficient, $update$ the parameter update strategy and f the nonlinear function. Note that [200, 50] means that *CDRR* uses two layers with the sizes of 200 and 50, respectively. And the learning rate for training *DSWE* is decayed by a factor of 0.8 per epoch.

Following Liu et al. (2016), we perform a 4-way classification on the four top-level relations in the PDTB: $Temporal$ ($Temp$), $Comparison$ ($Comp$), $Contingency$ ($Cont$) and $Expansion$ ($Expa$). The PDTB is split into the training set (Sections 2-20), development set (Sections 0-1) and test set (Sections 21-22). Table 2 lists the statistics of these data sets. Due to the small and uneven test data set, we run our method 10 times with different random seeds (therefore different initial parameters), and report the results (of a run) which are closest to the average results. Finally, we use both $Accuracy$ and $Macro\ F_1$ (macro-averaged F_1) to evaluate our method.

Relation	Train	Dev	Test
$Temp$	582	48	55
$Comp$	1855	189	145
$Cont$	3235	281	273
$Expa$	6673	638	538

Table 2: Statistics of data sets on the PDTB.

3.2 Results

We compare our learned discourse-specific word embeddings (*DSWE*) with two publicly available embeddings[3]:

1) *GloVe*[4]: trained on 6B words from *Wikipedia 2014* and *Gigaword 5* using the count based model in (Pennington et al., 2014), with a vocabulary of 400K and a dimensionality of 300.

2) *word2vec*[5]: trained on 100B words from *Google News* using the CBOW model in (Mikolov et al., 2013), with a vocabulary of 3M and a dimensionality of 300.

Results in Table 3 show that using *DSWE* gains significant improvements (one-tailed t-test with p<0.05) over using *GloVe* or *word2vec*, on both $Accuracy$ and $Macro\ F_1$. Furthermore, using *DSWE* achieves better performance across all relations on the F_1 score, especially for minority relations ($Temp$, $Comp$ and $Cont$). Overall, our *DSWE* can effectively incorporate discourse infor-

[3] The reasons for using those publicly available word embeddings are: 1) They are both trained on massive data. 2) It will be convenient for other people to reproduce our experiments. 3) Using *GloVe* or *word2vec* word embeddings trained on the same corpus as *DSWE* achieves worse performance than using these two public ones.

[4] http://nlp.stanford.edu/projects/glove/glove.6B.zip

[5] https://code.google.com/archive/p/word2vec/GoogleNews-vectors-negative300.bin.gz

CDRR		*+GloVe*	*+word2vec*	*+DSWE*
$Temp$	P	36.00	27.03	31.58
	R	16.36	18.18	21.82
	F_1	22.50	21.74	25.81
$Comp$	P	53.97	50.00	43.00
	R	23.45	20.00	29.66
	F_1	32.69	28.57	35.10
$Cont$	P	44.90	51.81	55.29
	R	40.29	36.63	42.12
	F_1	42.47	42.92	47.82
$Expa$	P	60.47	60.72	63.91
	R	76.21	81.60	79.00
	F_1	67.43	69.63	70.66
$Accuracy$		55.68	57.17	**58.85**
$Macro\ F_1$		41.27	40.71	**44.84**

Table 3: Results of using different word embeddings. We also list the Precision, Recall and F_1 score for each relation.

mation in explicit data, and thus benefits implicit discourse relation recognition.

We also compare our method with three recent systems which also use explicit data to boost the performance:

1) *R&X2015*: Rutherford and Xue (2015) construct weakly labeled data from explicit data based on the chosen connectives, to enlarge the training data directly.

2) *B&D2016*: Braud and Denis (2016) learn connective-based word representations and build a logistic regression model based on them[6].

3) *Liu2016*: Liu et al. (2016) use a multi-task neural network to incorporate several discourse-related data, including explicit data and the RST-DT corpus (William and Thompson, 1988).

System	$Accuracy$	$Macro\ F_1$
R&X2015	57.10	40.50
B&D2016	52.81	42.27
Liu2016	57.27	**44.98**
CDRR+DSWE	**58.85**	44.84

Table 4: Comparison with recent systems.

Results in Table 4 show the superiority of our method. Although *Liu2016* performs slightly better on $Macro\ F_1$, it uses the additional labeled RST-DT corpus. For *R&X2015* and *Liu2016*, they both incorporate relatively small explicit data because of the noise problem, for example, 20,000 and 40,000 instances respectively. By contrast, our method benefits from about 4.9M explicit instances. While *B&D2016* uses massive explicit data, it is limited by the fact that the maximum dimension of word representations is restricted to the number of connectives, for example 96 in their work. Overall, our method can effectively utilize massive explicit data, and thus is more powerful than baselines.

not		*good*	
word2vec	*DSWE*	*word2vec*	*DSWE*
do	no	great	great
did	n't	bad	lot
anymore	never	terrific	very
necessarily	nothing	decent	better
anything	neither	nice	success
anyway	none	excellent	well
does	difficult	fantastic	happy
never	nor	better	certainly
want	refused	solid	respect
neither	impossible	lousy	fine
if	limited	wonderful	import
know	declined	terrible	positive
anybody	nobody	Good	help
yet	little	tough	useful
either	denied	best	welcome

Table 5: Top 15 closest words of *not* and *good* in both *word2vec* and *DSWE*.

To give an intuition of what information is encoded into the learned *DSWE*, we list in Table 5 the top 15 closest words of *not* and *good*, according to the cosine similarity. We can find that, in *DSWE*, words similar to *not* to some extent have negative meanings. And since *declined* is similar to *not*, a classifier may easily identify the implicit instance *[A network spokesman would* ***not*** *comment. ABC Sports officials* ***declined*** *to be interviewed.]* as the *Conjunction* relation. For *good* in *DSWE*, the similar words no longer include words like *bad*. Furthermore, the similar score between *good* and *great* is 0.54 while the score between *good* and *bad* is just 0.33, which may make a classifier easier to distinguish word pairs *(good, great)* and *(good, bad)*, and thus is helpful for predicting the *Conjunction* relation. This qualitative analysis demonstrates the ability of our *DSWE* to capture the discourse relationships between words.

[6] We carefully reproduce their model since they adopt a different setting in preprocessing the PDTB.

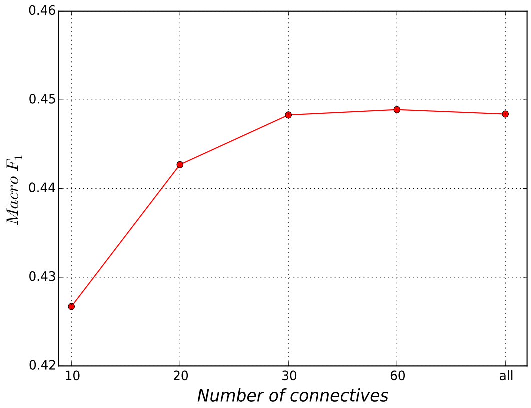

Figure 2: Impact of connectives used in training *DSWE*.

Finally, we conduct experiments to investigate the impact of connectives used in training *DSWE* on our results. Specifically, we use the explicit discourse instances with the top 10, 20, 30, 60 most frequent or all connectives to learn *DSWE*, accounting for 78.9%, 91.9%, 95.8%, 99.4% or 100% of total instances, respectively. The top 10 most frequent connectives are: *and*, *but*, *also*, *while*, *as*, *when*, *after*, *if*, *however* and *because*, which cover all four top-level relations defined in the PDTB. As illustrated in Figure 2, with only the top 10 connectives, the learned *DSWE* achieves better performance than the common word embeddings. We observe a significant improvement when using top 20 connectives, almost the best performance with top 30 connectives, and no further substantial improvement with more connectives. These results indicate that we can use only top n most frequent connectives to collect explicit discourse data for *DSWE*, which is very convenient for most languages.

4 Conclusion

In this paper, we learn discourse-specific word embeddings from massive explicit data for implicit discourse relation recognition. Experiments on the PDTB show that using the learned word embeddings as features can significantly boost the performance. We also show that our method can use explicit data more effectively than previous work. Since most of neural network models for implicit discourse relation recognition use pretrained word embeddings as input, we hope that our learned word embeddings would benefit them.

Acknowledgments

We would like to thank all the reviewers for their constructive and helpful suggestions on this paper. This work is partially supported by the Natural Science Foundation of China (Grant Nos. 61573294, 61672440, 61075058), the Ph.D. Programs Foundation of Ministry of Education of China (Grant No. 20130121110040), the Foundation of the State Language Commission of China (Grant No. WT135-10), the Natural Science Foundation of Fujian Province (Grant No. 2016J05161).

References

Or Biran and Kathleen McKeown. 2013. Aggregated Word Pair Features for Implicit Discourse Relation Disambiguation. In *Proceedings of ACL*. Sofia, Bulgaria, pages 69–73.

Chloé Braud and Pascal Denis. 2015. Comparing Word Representations for Implicit Discourse Relation Classification. In *Proceedings of EMNLP*. Lisbon, Portugal, pages 2201–2211.

Chloé Braud and Pascal Denis. 2016. Learning Connective-based Word Representations for Implicit Discourse Relation Identification. In *Proceedings of EMNLP*. Austin, Texas, pages 203–213.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network. In *Proceedings of ACL*. Berlin, Germany, pages 1726–1735.

Yangfeng Ji and Jacob Eisenstein. 2015. One Vector is Not Enough: Entity-Augmented Distributed Semantics for Discourse Relations. *Transactions of the Association for Computational Linguistics* 3:329–344.

Man Lan, Yu Xu, and Zhengyu Niu. 2013. Leveraging Synthetic Discourse Data via Multi-task Learning for Implicit Discourse Relation Recognition. In *Proceedings of ACL*. Sofia, Bulgaria, pages 476–485.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of EMNLP*. PA, USA, pages 343–351.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled End-to-end Discourse Parser. *Natural Language Engineering* 20(02):151–184.

Yang Liu and Sujian Li. 2016. Recognizing Implicit Discourse Relations via Repeated Reading: Neural Networks with Multi-Level Attention. In *Proceedings of EMNLP*. Austin, Texas, pages 1224–1233.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit Discourse Relation Classification via Multi-Task Neural Networks. In *Proceedings of AAAI*. Arizona, USA, pages 2750–2756.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using Entity Features to Classify Implicit Discourse Relations. In *Proceedings of SIGDIAL*. PA, USA, pages 59–62.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]* .

H. Peng, Fulmi Long, and C. Ding. 2005. Feature Selection Based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy. *IEEE Transactions on PAMI* 27(8):1226–1238.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP*. Doha, Qatar, pages 1532–1543.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic Sense Prediction for Implicit Discourse Relations in Text. In *Proceedings of ACL-IJCNLP*. PA, USA, pages 683–691.

Emily Pitler and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *Proceedings of ACL-IJCNLP*. PA, USA, pages 13–16.

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 LREC*. volume 24, pages 2961–2968.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A Stacking Gated Neural Architecture for Implicit Discourse Relation Classification. In *Proceedings of EMNLP*. Austin, Texas, pages 2263–2270.

Attapol Rutherford and Nianwen Xue. 2014. Discovering Implicit Discourse Relations Through Brown Cluster Pair Representation and Coreference Patterns. In *Proceedings of EACL*. Gothenburg, Sweden, pages 645–654.

Attapol Rutherford and Nianwen Xue. 2015. Improving the Inference of Implicit Discourse Relations via Classifying Explicit Discourse Connectives. In *Proceedings of NAACL*. Denver, Colorado, pages 799–808.

Mann William and Sandra Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text* 8(3):243–281.

Changxing Wu, Xiaodong Shi, Yidong Chen, Yanzhou Huang, and Jinsong Su. 2016. Bilingually-constrained Synthetic Data for Implicit Discourse Relation Recognition. In *Proceedings of EMNLP*. Austin, Texas, pages 2306–2312.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition. In *Proceedings of EMNLP*. Lisbon, Portugal, pages 2230–2235.