Automatic Prediction of Discourse Connectives

Eric Malmi[1,*], **Daniele Pighin**[2], **Sebastian Krause**[2,*], **Mikhail Kozhevnikov**[2]
[1]Aalto University
Espoo, Finland
eric.malmi@aalto.fi
[2]Google
Zürich, Switzerland
{biondo,bastik,qnan}@google.com

Abstract

Accurate prediction of suitable discourse connectives (*however*, *furthermore*, etc.) is a key component of any system aimed at building coherent and fluent discourses from shorter sentences and passages. As an example, a dialog system might assemble a long and informative answer by sampling passages extracted from different documents retrieved from the Web. We formulate the task of discourse connective prediction and release a dataset of 2.9M sentence pairs separated by discourse connectives for this task. Then, we evaluate the hardness of the task for human raters, apply a recently proposed decomposable attention (DA) model to this task and observe that the automatic predictor has a higher F_1 than human raters (32 vs. 30). Nevertheless, under specific conditions the raters still outperform the DA model, suggesting that there is headroom for future improvements.

Keywords: discourse connectives, decomposable attention model, discourse relation prediction

1. Introduction

Discourse connectives, also referred to as discourse markers, discourse cues, or discourse adverbials, are used to bind together and to explicate the relation between pieces of text. It is a common language class exercise to be asked to fill in suitable connectives to a text in order to improve the text flow. Similarly, it is important for computational summarization and text adaptation systems to be able to fill in suitable discourse connectives to produce natural-sounding utterances.

In this work, we study the problem of automatic discourse connective prediction. We limit ourselves to connectives which appear at the beginning of a sentence, linking the sentence to the preceding one. Even in this limited setting, an automatic discourse connective predictor has many concrete use cases. For example, in a question-answering setting it could help to generate answers by collating sentences from multiple sources. In extractive text summarization, it could be used to determine what is the best way to join two sentences that used to be separated by one or more sentences. As part of a text-authoring application, it could suggest suitable connectives at the beginning of a sentence. In the literature, discourse connective prediction has been recently studied merely as an intermediate step for the well-studied problem of implicit discourse relation prediction (Xu et al., 2012; Zhou et al., 2010). However, considering the aforementioned applications, we argue that connective prediction makes an interesting and relevant problem in its own right.

The contributions of this work are twofold:

1. We present an extensive experimental study on the problem of discourse connective prediction and show that a recently proposed decomposable attention model (Parikh et al., 2016) yields a good performance on this task. The model clearly outperforms a popular word-pair model and obtains a better performance than human raters on the same task and data.

2. We describe the dataset that we collected, consisting of 2.9 million adjacent sentence pairs (with and without a connective) extracted from the English Wikipedia. For 10 000 sentences, we also include connectives filled in by human raters. The dataset is publicly available at: https://github.com/ekQ/discourse-connectives

[*]Work performed during an internship at Google.

2. Related Work

A few earlier works study discourse connective prediction alone, but recently it has been studied merely as an intermediate step for discourse relation prediction. Next we provide a brief overview of these two lines of work, starting from the latter.

2.1. Predicting Connectives for Implicit Discourse Relation Prediction

Implicit discourse relation prediction has attracted considerable attention in recent years (Braud and Denis, 2016; Liu and Li, 2016; Qin et al., 2016; Qin et al., 2017; Rutherford and Xue, 2015; Wu et al., 2017; Zhang et al., 2016). Earlier Pitler et al. (2008) showed that if a discourse connective is known, the explicit discourse relation[1] can be inferred with a 93.09% accuracy, which has inspired several efforts at predicting connectives to improve implicit discourse relation prediction. Zhou et al. (2010) predicted connectives using an N-gram language model, whereas Xu et al. (2012) employed word pairs and a selection of linguistically informed features. Liu et al. (2016), on the other hand, showed that predicting both connectives and relations using a convolutional neural network in a multi-task setting improves the relation prediction performance.

2.2. Discourse Connective Prediction

Some earlier works have focused on connective prediction alone and developed various hand-crafted features for distinguishing between connectives. For example, Elhadad and McKeown (1990) explored pragmatic features for distinguishing between the connectives `but` and `although`,

[1]Later, it has been shown that a single discourse connective can actually convey multiple discourse relations (Rohde et al., 2015; Rohde et al., 2016).

and between because and since. Later Grote et al. (1998) developed a specialized lexicon for discourse connectives based on the relevant constraints and preferences associated with the connectives. While these works do not present an experimental evaluation of the proposed systems, we evaluate our connective prediction models extensively in order to understand their applicability to real-life scenarios. Furthermore, our aim is to learn the representations of the two arguments and their relationship automatically which allows us to distinguish between a large set of connectives without extensive manual efforts required to craft features that separate the connectives.

In addition to predicting the most suitable discourse connective, several methods have been developed for predicting the presence of a discourse connective (Yung et al., 2017; Di Eugenio et al., 1997; Patterson and Kehler, 2013). We also predict the presence of a connective by considering [No connective] as one of the classes to be predicted.

3. Data Collection

We compile a list of 79 discourse connectives based on the Penn Discourse Treebank (PDTB) (Prasad et al., 2008). Since our focus is on sentence concatenation, we ignore the (forward) connectives, which typically point to the following sentence rather than the previous one, such as "After the election, [...]". However, for several ambiguous connectives, the forward use can be ruled out by requiring a comma after the connective (e.g. Instead,); we include such connectives in our data. Discontinuous connectives, such as "If [...] then [...]", are not included.

Data samples for discourse connective prediction can be collected from any large unannotated text corpus. In this instance, we use the English Wikipedia[2] and collect every pair of consecutive sentences within the same paragraph where the latter sentence begins with one of the 79 discourse connectives. As a result, we obtain a dataset of 1.95 million sentence pairs separated by a connective. Additionally, we collect 0.91 million examples of consecutive sentences not separated by a discourse connective, labeled as [No connective], for a total of 2.86 million sentence pairs.[3]

The frequency distribution of the connectives is very skewed; however occurs 720 334 times, whereas else, only 43 times in the beginning of a sentence. In order to make the connective prediction task more feasible for the models and for human raters, we select a subset of sufficiently frequent and distinct connectives (e.g. for example is included but for instance is not since it conveys the same meaning and is less frequent). The details of the selection process are omitted in the interest of space, but the resulting 19 connectives are listed in Table 1.

Finally, we split the data into train, development, and test sets. We balance the connective classes, since in an unbalanced dataset most examples would be labeled as [No connective] and many connectives would be extremely under-represented, limiting the applicability of the resulting classifier. For the development and test sets, we pick 500 samples per connective (including [No connective]) by under-sampling without replacement. This results in two balanced datasets of 10 000 samples. For the training set, we pick 20 000 samples per connective by under-sampling the majority classes and oversampling the minority classes, creating a balanced dataset of 400 000 samples. Connective samples from a single Wikipedia article are not included in more than one of the three datasets to avoid over-fitting through potential repetition within a single article.

In comparison with the PDTB dataset, which contains information about both discourse connectives and discourse relations, the main advantage of the collected dataset is its size. PDTB contains only 40 600 examples (1.4% of the size of the collected dataset), which causes sparsity issues (Li and Nenkova, 2014). This can slow down the development of new models, particularly complex neural models that often require large training datasets to generalize well.

Connective	Occurrences	Connective	Occurrences
however	720 334	on the other hand	20 301
for example	111 711	in particular,	16 011
and	73 644	indeed,	15 286
meanwhile,	57 971	overall,	9 513
therefore	44 064	in other words	8 888
finally,	33 076	rather,	5 596
nevertheless	32 952	by contrast,	4 605
instead,	30 973	by then	4 279
moreover	25 583	otherwise,	3 563
then,	21 731		

Table 1: The list of 19 connectives studied in the experiments in addition to the [No connective] class. Only the connectives which are sufficiently frequent and distinct in their meaning have been selected.

4. Connective Prediction Models

The decomposable attention (DA) model was recently introduced by Parikh et al. (2016) for the *natural language inference* (NLI) problem which aims to classify entailment and contradiction relations between a premise and a hypothesis. Discourse connective prediction is related to the NLI problem since entailment and contradiction can be explicitly indicated by certain connectives (for instance, therefore and by contrast, respectively). However, the larger number of classes makes connective prediction more challenging. DA was shown to yield a state-of-the-art performance on the NLI task while requiring almost an order of magnitude fewer parameters than previous approaches. For all these reasons, it seems natural to apply the DA model to the connective prediction problem.

Marcu and Echihabi (2002) proposed to use word-pair features to predict discourse relations based on discourse connectives mapped to these relations. Similarly, many later implicit discourse relation prediction models are based on word-pair features (Marcu and Echihabi, 2002; Pitler et al., 2009; Xu et al., 2012; Zhou et al., 2010) or aggregated word-pair features (Biran and McKeown, 2013; Rutherford and Xue, 2014). Therefore, we use a model called WORDPAIRS to have a baseline for the DA model.

[2] A snapshot from September 5, 2016.

[3] Note that our models are tested only on consecutive sentences, for which the ground truth connectives are known, but they can be applied to connect also disjoint sentences.

4.1. The Decomposable Attention Model

The DA model consists of three steps, *attend*, *compare*, and *aggregate*, which are executed by three different feed-forward neural networks F, G, and H, respectively. As input, the model takes two sentences $\mathbf{a}$ and $\mathbf{b}$ represented by sequences of word embeddings. The sequences are padded by "NULL" tokens to fix their lengths to 50 tokens.

In the **attend** step, the model computes non-negative attention scores for each pair of tokens across the two input sentences. This computation ignores the order of the tokens and it produces soft-alignments from $\mathbf{a}$ to $\mathbf{b}$ and *vice versa*.

In the **compare** step, the model computes comparison vectors between each input token and its aligned sub-phrase. The aligned sub-phrase is a linear combination of the embedding vectors of the other sentence weighted by the attention scores.

Finally, in the **aggregate** step, the comparison vectors are summed over the tokens of a sentence and then the aggregate vectors of the two sentences are concatenated. The resulting vector is fed into the third feed-forward network which outputs $\hat{\mathbf{y}}$ containing scores for each class. The predicted class is given by $\hat{y} = \arg\max_i \hat{y}_i$.

The weights of the three networks are randomly initialized, after which the model is trained in an end-to-end manner. Our implementation of the DA model has the following differences compared to the original model described by Parikh et al. (2016): (i) we do not use the self-attention mechanism which was reported to provide only a small improvement over the vanilla version of DA; (ii) we do not project down the embedding vectors but use 100-dimensional word2vec embeddings (Mikolov et al., 2013) which are updated during the training; (iii) we use layer normalization (Ba et al., 2016) which makes the model converge faster.

4.2. The Word-Pair Model

The WORDPAIRS model considers as features all word pairs which appear across the two arguments (e.g. word A appears in Arg 1 and word B in Arg 2) in at least five samples in the training dataset. Such features are employed by many implicit discourse relation prediction models (Marcu and Echihabi, 2002; Pitler et al., 2009; Zhou et al., 2010; Xu et al., 2012). Additionally, we incorporate single word features (e.g. word *A* appears in Arg 2) since these slightly improved the results. With these binary features, we train logistic regressors using the one-vs-rest scheme to predict one of the 20 different connectives.[4]

5. Experiments

Next we present experimental results on discourse connective prediction using human raters, the DA model and the WORDPAIRS model. For this task, we remove the connective (if any) from the second sentence in each test pair, and measure the ability of the model (or the raters) to identify the removed connective.

[4]We trained two versions of the WORDPAIRS model: using stochastic gradient descent with mini-batches and using LIBLINEAR with 100k samples (i.e. 25% of the training data) which we could fit into the memory of a 256 GB machine. The reported results are based on the latter approach, which performed better.

Model	Macro F_1	Accuracy
RANDOM	5.00	5.00
WORDPAIRS	14.81	15.60
DA	**31.80**	**32.71**
Human Raters	23.72	23.12

Table 2: Discourse connective prediction performance of a RANDOM baseline, the WORDPAIRS model, the decomposable attention model (DA) and human raters.

5.1. Accuracy of Human Raters

To better understand what is reasonable to expect from an automatic predictor, we use a crowd-sourcing platform to ask human raters to reconstruct the removed connectives for each of the 10 000 test sentence pairs. Each sentence pair is annotated by three (not necessarily the same three) native English speakers. The raters are shown the two sentences, the latter of which starts with a *[Connective goes here]* placeholder, and asked to select the most suitable connective from the 20 options, including `[No connective]`. This layer of human annotations is also released as part of the connective dataset. The raters are instructed to pick the most natural connective in case there are multiple suitable options. Furthermore, they are asked to pick `[No connective]` only if adding a connective would make the concatenation sound ungrammatical or artificial, or if the two sentences seem to be completely disconnected. The sentences are not pre-processed apart from upper-casing the first character of the second sentence to avoid giving away the presence of a connective in the original sentence. The order of the connectives is randomized, except for `[No connective]` which is always shown last.

On the whole test-set, human annotators achieve a macro-averaged F_1 score of 23.72. The confusion matrix generated by the raters' decisions is presented on the left side of Figure 1. It shows that the raters are strongly biased towards `[No connective]` despite the indication to refrain from using it. A similar bias was observed by Rohde et al. (2016) for the task of filling in a suitable conjunction before a discourse connective. There are at least two possible explanations for this bias: (i) in the sake of clarity and in line with common scientific writing guidelines, Wikipedia editors tend to use connectives quite generously, and (ii) the artificial balancing of the datasets makes `[No connective]` under-represented in the test data compared to the actual distribution of discourse connectives vs. `[No connective]`. The confusion matrix also shows that there are clusters of connectives that raters tend to confuse, even though they do not necessarily encode exactly the same relation. Examples are `rather,` and `instead,`, `for example` and `in particular,`, `on the other hand` and `by contrast,`. For 57.1% of the test questions, there is a consensus among at least two raters and for 11.4%, all three raters agree on the most suitable connective.

5.2. Accuracy of the Models

In this section, the DA model and the WORDPAIRS model are employed to perform the same task as the human raters,

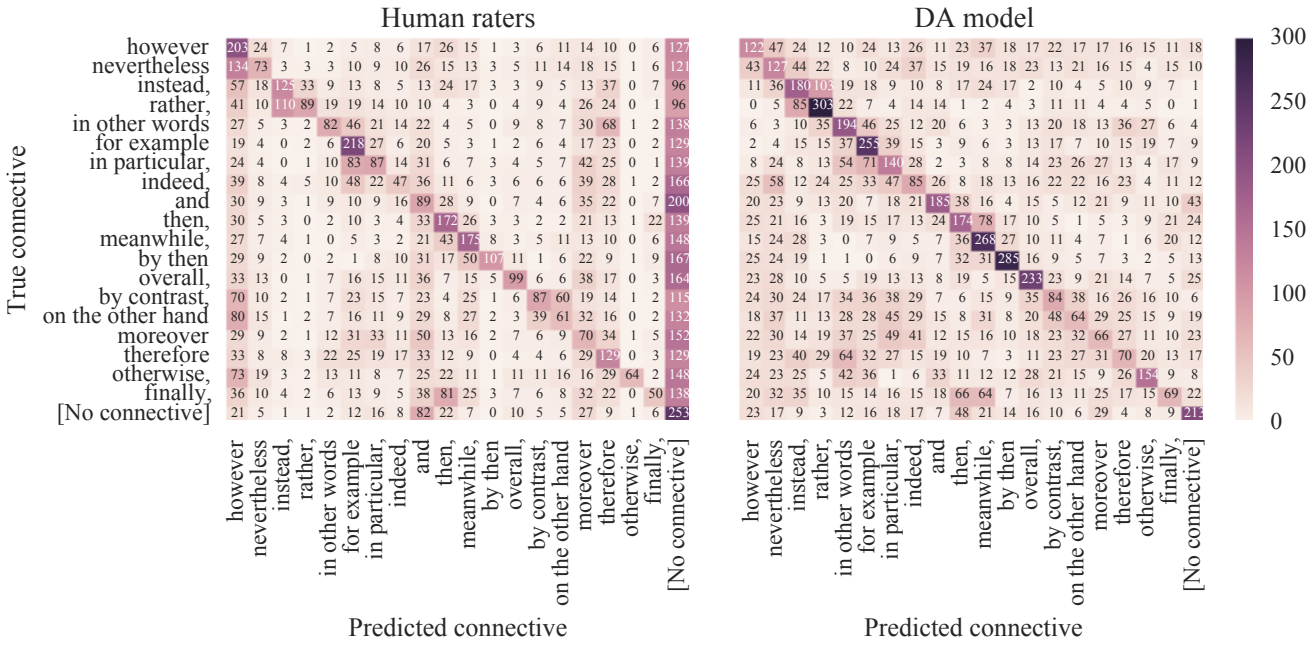

Figure 1: Confusion matrices for the human raters (left) and DA model (right) predicting the discourse connectives used by Wikipedia authors. The counts of the human raters are divided by three (i.e., the number of raters) for easier comparison.

i.e., learning to reconstruct the connective possibly removed from the beginning of the second sentence in each test pair. A balanced dataset is used for both training and testing the models as described in Section 3. The DA model is evaluated using the following hyper-parameters optimized on the development set: network size (one hidden layer with 200 neurons), batch size (64), dropout ratios for the F, G, and H networks (0.68, 0.14, and 0.44, respectively), and learning rate (0.0018). The model is implemented in TensorFlow (Abadi et al., 2015) and the training is run for 300 000 batch steps. The results, reported in Table 2, show that DA clearly outperforms the WORDPAIRS baseline with an F_1 score of 31.80 vs. 14.81.

5.3. Comparison Between the Raters and the Decomposable Attention Model

Table 3 compares the accuracy of DA predictions to the rater decisions. The macro-averaged F_1 score of human raters is 23.72 which is, quite surprisingly, lower than the F_1 score of the DA model, 31.80. The difference is smaller when considering majority votes on the subset of 5 714 tasks for which there is a consensus among at least 2 out of 3 raters, which results in a 30.36 F_1 score for the raters. On these less ambiguous cases, the model performance also increases to 32.68.

As we mentioned in Section 5.1., human raters are clearly less eager to introduce a connective than Wikipedia editors. Therefore, we also evaluate the setting in which we exclude the questions for which either the ground-truth label, or the rater-assigned majority label, or the model-assigned label is `[No connective]`. The results, listed in the last line of Table 3, show that under these conditions human raters actually outperform the model.

The confusion matrix of DA is shown on the right side of Figure 1. For each connective, the true connective is the most frequent prediction. Connective `on the other hand` has the lowest F_1 score (15.06), whereas `by then` has the highest (57.29). Some of the most frequent mistakes are between similar connectives, such as `however` vs. `nevertheless`, and `instead,` vs. `rather,`. These errors are by and large consistent with those of human raters (left side of the figure). This confirms that the model is accurately capturing the meaning of the relation, and when it does not select the gold connective it is making similar approximations to what people would do. Furthermore, the Figure 1 shows that raters have a more pronounced tendency to select frequent connectives, such as `however` and `and`. To further exemplify, in Table 4 we show a selection of wrong and correct decisions made by DA and human raters. A manual inspection of these and other examples shows that in some cases a larger context than the previous sentence is required for inferring the connective. For instance, to correctly decide whether `finally,` is more suitable than `then,` one may have to inspect a larger context.

Setting	n	Raters (F_1)	Model (F_1)
A	10 000	23.72	**31.80**
B	5 714	30.36	**32.68**
C	3 204	**41.97**	36.65

Table 3: Macro-averaged F_1 scores for human raters and the decomposable attention model. The three settings are: (A) All test set items; (B) Only the items for which there is a consensus among at least 2 out of 3 raters; (C) Consensus items ignoring those where either ground truth, or the rater assigned, or the model assigned label is `[No connective]`.

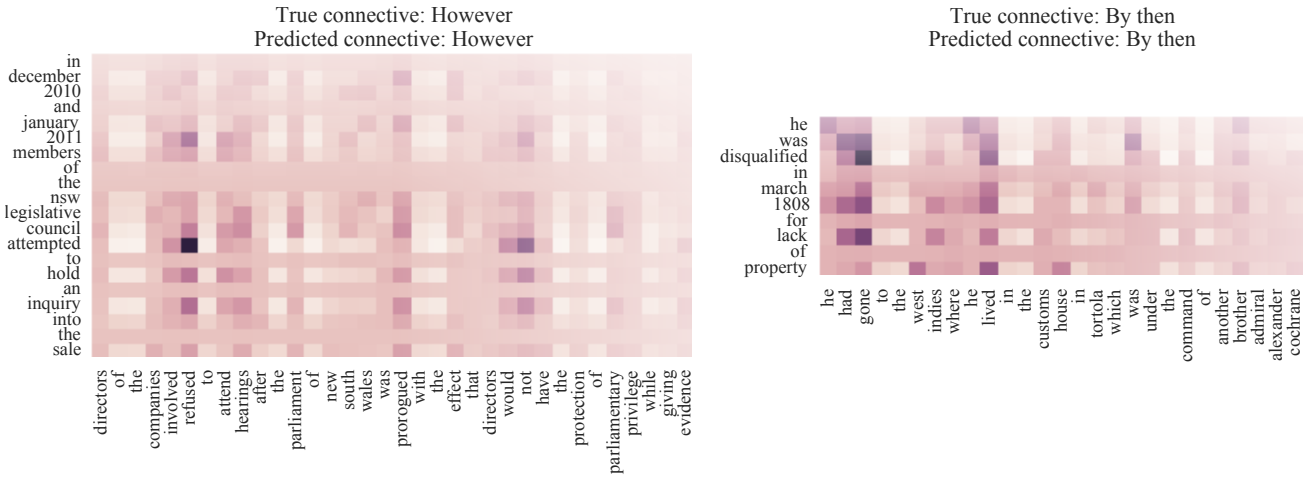

Figure 2: Two examples of the alignment matrix between the first (y-axis) and the second argument (x-axis) generated by the DA predictor. The darker the color, the higher the alignment score.

Arguments	Gold	DA	Raters
Arg 1: From 1913 to 1917 Lucey served as Illinois Attorney General. Arg 2: Lucey was appointed to the Illinois Public Utilities Commission in 1917 and served until 1920.	**then**	**then**	[No connective]
Arg 1: In Ahmedgarh Municipal Council, Female Sex Ratio is of 901 against state average of 895. Arg 2: Child Sex Ratio in Ahmedgarh is around 841 compared to Punjab state average of 846.	**moreover**	**moreover**	meanwhile
Arg 1: If the question was answered correctly, the team would receive the next clue. Arg 2: The chosen team member would have to try again.	**otherwise**	**otherwise**	then
Arg 1: Lee then continued on to Boston, arriving 25 June. Arg 2: The ranks of General Washington's Navy were being thinned by captures.	**meanwhile**	by then	**meanwhile**
Arg 1: Cooper was promoted as an alternate leader to Ahern. Arg 2: It was thought he could shore up the National Party's vote in its conservative rural heartland.	**in particular**	however	**in particular**
Arg 1: Taylor urged Pius XII to explicitly condemn Nazi atrocities. Arg 2: Pius XII spoke against the "evils of modern warfare", but did not go further.	**instead**	in particular	**instead**

Table 4: Examples of mistakes and correct predictions made by the DA model and by the raters.

5.4. Model Interpretation

An advantage of the DA model is that it is possible to examine which words the model attends to when inferring a connective. In some cases, the attended words are clearly meaningful semantically or linguistically, whereas in other cases the soft-alignment matrix that the model produces is harder to interpret. Examples of the former case are represented in Figure 2, which shows the alignment matrices from the tokens of the first sentence (y-axis) to the tokens of the second sentence (x-axis) so that the rows sum to 1. In the left example, the model correctly predicts `however` as the connective after aligning the word *attempt* with *refuse* and *not*. These word pairs indicate contrast which makes `however` a likely connective. In the right example, the model aligns the phrase *was disqualified* with *had gone* and correctly predicts `by then` as the connective. The corresponding tenses, i.e., past and past perfect, respectively, are likely clues of the presence of `by then`.

6. Conclusions

We studied the problem of discourse connective prediction, which has many useful applications in text summarization, adaptation and conversationalization. We collected a dataset of 2.9 million pairs of consecutive sentences and connectives, and made it publicly available to facilitate further research on this problem, as well as other related bi-sequence classification tasks. We showed that the recently proposed decomposable attention model performs surprisingly well on the connective prediction task, even better than human raters on the same representative test set consisting of 10 000 samples. We also observed that, unlike the model, human raters have a preference for implicit connectives, as they do outperform the model if the comparison is restricted to the cases in which the majority of raters agrees on an explicit connective. The alignment matrices produced by the model suggest that the predictor is picking up relevant lexical, syntactic and semantic clues. The confusion matrix of the predictor shows very similar error patterns to the matrix generated from human raters, further confirming the meaningfulness of the decisions made by the model.

7. Acknowledgements

We would like to thank Cesar Ilharco for his help on running the experiments.

8. Bibliographical References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2015). Tensorflow: Large-scale machine learning on heterogeneous systems. *Software available from tensorflow.org*.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Biran, O. and McKeown, K. (2013). Aggregated word pair features for implicit discourse relation disambiguation. In *Proc. ACL*.

Braud, C. and Denis, P. (2016). Learning connective-based word representations for implicit discourse relation identification. In *Proc. EMNLP*.

Di Eugenio, B., Moore, J. D., and Paolucci, M. (1997). Learning features that predict cue usage. In *Proc. EACL*.

Elhadad, M. and McKeown, K. R. (1990). Generating connectives. In *Proc. the 13th Conference on Computational Linguistics*.

Grote, B., Stede, M., et al. (1998). Discourse marker choice in sentence planning. In *Proc. Ninth International Workshop on Natural Language Generation*.

Li, J. J. and Nenkova, A. (2014). Reducing sparsity improves the recognition of implicit discourse relations. In *Proc. SIGDIAL*.

Liu, Y. and Li, S. (2016). Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proc. EMNLP*.

Liu, Y., Li, S., Zhang, X., and Sui, Z. (2016). Implicit discourse relation classification via multi-task neural networks. In *Proc. AAAI*.

Marcu, D. and Echihabi, A. (2002). An unsupervised approach to recognizing discourse relations. In *Proc. ACL*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*.

Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. (2016). A decomposable attention model for natural language inference. In *Proc. EMNLP*.

Patterson, G. and Kehler, A. (2013). Predicting the presence of discourse connectives. In *Proc. EMNLP*.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. K. (2008). Easily identifiable discourse relations. In *Proc. Coling 2008: Companion volume: Posters and Demonstrations*.

Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proc. ACL-IJCNLP*.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The penn discourse treebank 2.0. In *Proc. LREC*.

Qin, L., Zhang, Z., and Zhao, H. (2016). A stacking gated neural architecture for implicit discourse relation classification. In *Proc. EMNLP*.

Qin, L., Zhang, Z., Zhao, H., Hu, Z., and Xing, E. P. (2017). Adversarial connective-exploiting networks for implicit discourse relation classification.

Rohde, H., Dickinson, A., Clark, C., Louis, A., and Webber, B. (2015). Recovering discourse relations: Varying influence of discourse adverbials. In *Proc. LSDSem Workshop*.

Rohde, H., Dickinson, A., Schneider, N., Clark, C. N., Louis, A., and Webber, B. (2016). Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proc. LAW X Workshop*.

Rutherford, A. and Xue, N. (2014). Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proc. EACL*.

Rutherford, A. and Xue, N. (2015). Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proc. NAACL-HLT*.

Wu, C., Shi, X., Chen, Y., Huang, Y., and Su, J. (2017). Leveraging bilingually-constrained synthetic data via multi-task neural networks for implicit discourse relation recognition. *Neurocomputing*, 243:69–79.

Xu, Y., Lan, M., Lu, Y., Niu, Z. Y., and Tan, C. L. (2012). Connective prediction using machine learning for implicit discourse relation classification. In *Proc. IJCNN*.

Yung, F., Duh, K., Komura, T., and Matsumoto, Y. (2017). A psycholinguistic model for the marking of discourse relations. *Dialogue & Discourse*, 8(1):106–131.

Zhang, B., Xiong, D., Su, J., Liu, Q., Ji, R., Duan, H., and Zhang, M. (2016). Variational neural discourse relation recognizer. In *Proc. EMNLP*.

Zhou, Z. M., Lan, M., Niu, Z. Y., Xu, Y., and Su, J. (2010). The effects of discourse connectives prediction on implicit discourse relation recognition. In *Proc. SIGDIAL*.