Discourse Relation Sense Classification Using Cross-argument Semantic Similarity Based on Word Embeddings

Todor Mihaylov and Anette Frank
Research Training Group AIPHES
Department of Computational Linguistics
Heidelberg University, 69120 Heidelberg, Germany
{mihaylov,frank}@cl.uni-heidelberg.de

Abstract

This paper describes our system for the CoNLL 2016 Shared Task's supplementary task on Discourse Relation Sense Classification. Our official submission employs a Logistic Regression classifier with several cross-argument similarity features based on word embeddings and performs with overall F-scores of 64.13 for the *Dev* set, 63.31 for the *Test* set and 54.69 for the *Blind* set, ranking first in the *Overall* ranking for the task. We compare the feature-based Logistic Regression classifier to different Convolutional Neural Network architectures. After the official submission we enriched our model for Non-Explicit relations by including similarities of explicit connectives with the relation arguments, and part of speech similarities based on modal verbs. This improved our *Non-Explicit* result by 1.46 points on the *Dev* set and by 0.36 points on the *Blind* set.

1 Introduction

The CoNLL 2016 Shared Task on Shallow Discourse Parsing (Xue et al., 2016) focuses on identifying individual discourse relations presented in text. This year the shared task has a main track that requires end-to-end discourse relation parsing and a supplementary task that is restricted to discourse relation sense classification. For the main task, systems are required to build a system that given a raw text as input can identify arguments *Arg1* and *Arg2* that are related in the discourse, and also classify the type of the relation, which can be *Explicit*, *Implicit*, *AltLex* or *EntRel*. A further attribute to be detected is the relation *Sense*, which can be one of 15 classes organized hierarchically in 4 parent classes. With this work we participate in the Supplementary Task on Discourse Relation Sense Classification in English. The task is to predict the discourse relation sense when the arguments *Arg1*, *Arg2* are given, as well as the *Discourse Connective* in case of explicit marking.

In our contribution we compare different approaches including a Logistic Regression classifier using similarity features based on word embeddings, and two Convolutional Neural Network architectures. We show that an approach using only word embeddings retrieved from *word2vec* (Mikolov et al., 2013) and cross-argument similarity features is simple and fast, and yields results that rank first in the *Overall*, second in the *Explicit* and forth in the *Non-Explicit* sense classification task. Our system's code is publicly accessible[1].

2 Related Work

This year's CoNLL 2016 Shared Task on Shallow Discourse Parsing (Xue et al., 2016) is the second edition of the shared task after the CoNLL 2015 Shared task on Shallow Discourse Parsing (Xue et al., 2015). The difference to last year's task is that there is a new Supplementary Task on Discourse Relation Sense classification, where participants are not required to build an end-to-end discourse relation parser but can participate with a sense classification system only.

Discourse relations in the task are divided in two major types: Explicit and Non-Explicit (*Implicit*, *EntRel* and *AltLex*). Detecting the sense of Explicit relations is an easy task: given the discourse connective, the relation sense can be determined with very high accuracy (Pitler et al., 2008). A challenging task is to detect the sense of Non-Explicit discourse relations, as they usually don't

[1] https://github.com/tbmihailov/conll16st-hd-sdp - Source code for our Discourse Relation Sense Classification system

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 100–107,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

have a connective that can help to determine their sense. In last year's task *Non-Explicit* relations have been tackled with features based on Brown clusters (Chiarcos and Schenk, 2015; Wang and Lan, 2015; Stepanov et al., 2015), VerbNet classes (Kong et al., 2015; Lalitha Devi et al., 2015) and MPQA polarity lexicon (Wang and Lan, 2015; Lalitha Devi et al., 2015). Earlier work (Rutherford and Xue, 2014) employed Brown cluster and coreference patterns to identify senses of implicit discourse relations in naturally occurring text. More recently Rutherford and Xue (2015) improved inference of implicit discourse relations via classifying explicit discourse connectives, extending prior research (Marcu and Echihabi, 2002; Sporleder and Lascarides, 2008). Several neural network approaches have been proposed, e.g., Multi-task Neural Networks (Liu et al., 2016) and Shallow-Convolutional Neural Networks (Zhang et al., 2015). Braud and Denis (2015) compare word representations for implicit discourse relation classification and find that denser representations systematically outperform sparser ones.

3 Method

We divide the task into two subtasks, and develop separate classifiers for Explicit and Non-Explicit discourse relation sense classification, as shown in Figure 1. We do that because the official evaluation is divided into Explicit and Non-Explicit (Implicit, AltLex, EntRel) relations and we want to be able to tune our system accordingly. During training, the relation type is provided in the data, and samples are processed by the respective classifier models in *Process 1 (Non-Explicit)* and *Process 2 (Explicit)*. During testing the gold *Type* attribute is not provided, so we use a simple heuristic: we assume that *Explicit* relations have connectives and that *Non-Explicit*[2] relations do not.

As the task requires that the actual evaluation is executed on the provided server, we save the models so we can load them later during evaluation.

For classifying *Explicit* connectives we follow a feature-based approach, developing features based on word embeddings and semantic similarity measured between parts of the arguments *Arg1* and *Arg2* of the discourse relations. Classification is into one of the given fifteen classes of relation senses. For detecting *Non-Explicit* discourse relations we also make use of a feature-based approach, but in addition we experiment with two models based on Convolutional Neural Networks.

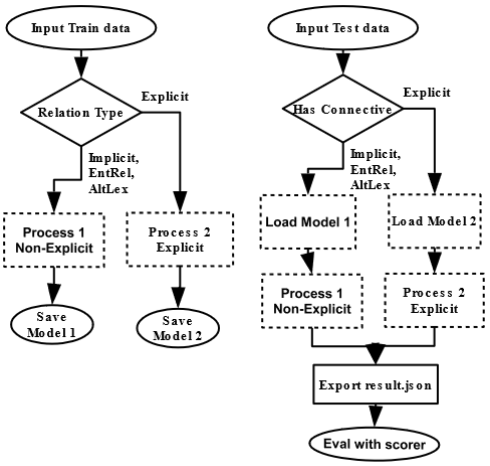

Figure 1: System architecture: Training and evaluating models for Explicit and Non-Explicit discourse relation sense classification

3.1 Feature-based approach

For each relation, we extract features from *Arg1*, *Arg2* and the *Connective*, in case the type of the relation is considered *Explicit*.

Semantic Features using Word Embeddings. In our models we only develop features based on word embedding vectors. We use *word2vec* (Mikolov et al., 2013) word embeddings with vector size 300 pre-trained on Google News texts.[3] For computing similarity between embedding representations, we employ cosine similarity:

$$1 - \frac{u.v}{\|u\| \, . \, \|v\|} \tag{1}$$

Embedding representations for Arguments and Connectives. For each argument *Arg1*, *Arg2* and *Connective* (for Explicit relations) we construct a centroid vector (2) from the embedding vectors $\vec{w_i}$ of all words w_i in their respective surface yield.

$$centroid(\vec{w}_1 ... \vec{w}_n) = \frac{\sum_{i=1}^{n} \vec{w}_i}{n} \tag{2}$$

[2] In fact, some *AltLex* discourse relations do have connectives, but they are considered *Non-Explicit*. More detailed analysis will be required to improve on this simple heuristic. Given that their distribution across the data sets is very small, they do not have much influence on the overall performance.

[3] `https://code.google.com/archive/p/word2vec/` - Pre-trained vectors trained on part of Google News dataset (about 100 billion words).

Cross-argument Semantic Vector Similarities. We calculate various similarity features on the basis of the centroid word vectors for the arguments and the connective, as well as on parts of the arguments:

Arg1 to Arg2 similarity. We assume that for given arguments *Arg1* and *Arg2* that stand in a specific discourse relation sense, their centroid vectors should stand in a specific similarity relation to each other. We thus use their cosine similarity as a feature.

Maximized similarity. Here we rank each word in *Arg2*'s text according to its similarity with the centroid vector of *Arg1*, and we compute the average similarity for the top-ranked N words. We chose the similarity scores of the top 1,2,3 and 5 words as features. The assumption is that the average similarity between the first argument (*Arg1*) and the top N most similar words in the second argument (*Arg2*) might imply a specific sense.

Aligned similarity. For each word in *Arg1*, we choose the most similar word from the yield of *Arg2* and we take the average of all best word pair similarities, as suggested in Tran et al. (2015).

Part of speech (POS) based word vector similarities. We used part of speech tags from the parsed input data provided by the organizers, and computed similarities between centroid vectors of words with a specific tag from *Arg1* and the centroid vector of *Arg2*. Extracted features for POS similarities are symmetric: for example we calculate the similarity between *Nouns* from *Arg1* with *Pronouns* from *Arg2* and the opposite. The assumption is that some parts of speech between *Arg1* and *Arg2* might be closer than other parts of speech depending on the relation sense.

Explicit discourse connectives similarity. We collected 103 explicit discourse connectives from the Penn Discourse Treebank (Prasad et al., 2008) annotation manual[4] and for all of them construct vector representations according to (2), where for multi-token connectives we calculate a centroid vector from all tokens in the connective. For every discourse connective vector representation we calculate the similarity with the centroid vector representations from all *Arg1* and *Arg2* tokens. This results in adding 103 similarity features for every relation. We use these features for implicit discourse relations sense classification only.

We assume that knowledge about the relation sense can be inferred by calculating the similarity between the semantic information of the relation arguments and specific discourse connectives.

Our feature-based approach yields very good results on Explicit relations sense classification with an F-score of 0.912 on the *Dev* set. Combining features based on word embeddings and similarity between arguments in Mihaylov and Nakov (2016) yielded state-of-the art performance in a similar task setup in Community Question Answering (Nakov et al., 2016), where two text arguments (question and answer) are to be ranked.

3.2 CNNs for sentence classification

We also experiment with Convolutional Neural Network architectures to detect Implicit relation senses. We have implemented the CNN model proposed in Kim (2014) as it proved successful in tasks like sentence classification and modal sense classification (Marasović and Frank, 2016). This model (Figure 2) defines one convolutional layer that uses pre-trained *Word2Vec* vectors trained on the Google News dataset. As shown in Kim (2014), this architecture yields very good results for various single sentence classification tasks. For our relation classification task we input the concatenated tokens of *Arg1* and *Arg2*.

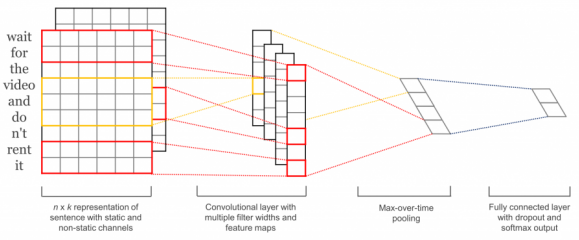

Figure 2: CNN architecture by Kim (2014).

3.3 Modified ARC-1 CNN for sentence matching

An alternative model we try for Implicit discourse relation sense classification is a modification of the *ARC-1* architecture proposed for sentence matching by Hu et al. (2015). We will refer to this model as *ARC-1M*. The modified architecture is depicted in Figure 3. The input of the model are two sentences S_x and S_y represented as sequence of

[4] https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf - The Penn Discourse Treebank 2.0 Annotation Manual

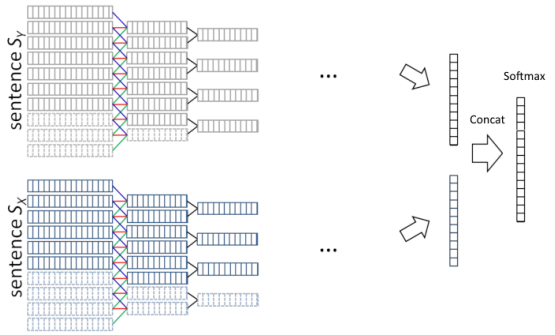

Figure 3: Modified ARC-I CNN architecture for sentence matching.

tokens' vector representations of *Arg1* and *Arg2*. Here, separate convolution and max-pooling layers are constructed for the two input sentences, and the results of the max-pooling layers are concatenated and fed to a single final *SoftMax* layer. The original ARC-1 architecture uses a *Multilayer Perceptron* layer instead of *SoftMax*. For our implementation we use TensorFlow (Abadi et al., 2015).

4 Experiments and Results

4.1 Data

In our experiments we use the official data (English) provided from the task organizers: *Train* (15500 Explicit + 18115 Non-Explicit), *Dev* (740 Explicit + 782 Non-Explicit), *Test* (990 Explicit + 1026 Non-Explicit), *Blind* (608 Explicit + 661 Non-Explicit). All models are trained on *Train* set.

4.2 Classifier settings

For our feature-based approach we concatenate the extracted features in a feature vector, scale their values to the 0 to 1 range, and feed the vectors to a classifier. We train and evaluate a L2-regularized Logistic Regression classifier with the LIBLINEAR (Fan et al., 2008) solver as implemented in *scikit-learn* (Pedregosa et al., 2011). For most of our experiments, we tuned the classifier with different values of the C (cost) parameter, and chose C=0.1 as it yielded the best accuracy on 5-fold cross-validation on the training set. We use these settings for all experiments that use the logistic regression classifier.

4.3 Official submission (LR with E+Sim)

Our official submission uses the feature-based approach described in Section 3.1 for both *Explicit* and *Non-Explicit* relations with all features described above, except for the *Explicit connective similarities (Conn)* and *Modal verbs similarities (POS MD)* which have been added after the submission deadline. Table 1 presents the results divided by senses from our official submission performed on the TIRA evaluation platform (Potthast et al., 2014) server. We also compare our official and improved system results to the best performing system in the CoNLL 2015 Shared Task (Wang and Lan, 2015) and the best performing systems in the CoNLL 2016 Discourse Relation Sense Classification task. With our official system we rank first in the *Overall*[5] ranking. We rank second in the *Explicit* ranking with a small difference of 0.07 behind the best system and fourth in the *Non-Explicit* ranking with more significant difference of 2.75 behind the best system. We can see that similar to (Wang and Lan, 2015) our system performs well in classifying both types, while this year's winning systems perform well in their winning relation type and much worse in the others[6].

4.4 Further experiments on Non-Explicit relations

In Table 2 we compare different models for *Non-Explicit* relation sense classification trained on the *Train* and evaluated on the *Dev* set.

Embeddings only experiments. The first three columns show the results obtained with three approaches that use only features based on word embeddings. We use *word2vec* word embeddings. We also experimented with pre-trained *dependency-based* word embeddings (Levy and Goldberg, 2014), but this yielded slightly worse results on the *Dev* set.

Logistic Regression (LR). The *LR* column shows the results from a Logistic Regression classifier that uses only the concatenated features from the centroid representations built from the words of *Arg1* and *Arg2*.

CNN experiments. The *CNN* column shows results obtained from the Convolutional Neural Network for sentence classification (Section 3.2) fed with the concatenated *Arg1* and *Arg2* word tokens' vector representations from *Word2Vec* word embeddings. For our experiments we used default

[5] *Overall* score is the F-score on All (both *Explicit* and *Non-Explicit*) relations.

[6] The winner team in *Non-Explicit* (Rutherford and Xue, 2016) does not participate in *Explicit*.

	WSJ Dev Set			WSJ Test Set			Blind Set (Official task ranking)		
Sense	**Overall**	**Exp**	**Non-E**	**Overall**	**Exp**	**Non-E**	**Overall**	**Exp**	**Non-E**
Comparison.Concession	33.33	40.00	0.00	36.36	44.44	0.00	91.67	100.00	0.00
Comparison.Contrast	74.31	94.44	16.07	65.99	92.19	9.60	21.24	25.81	0.00
Contingency.Cause.Reason	51.48	78.95	38.51	64.36	94.03	47.93	35.71	82.61	18.03
Contingency.Cause.Result	38.94	91.43	15.38	40.74	100.00	17.53	53.33	91.67	27.78
Contingency.Condition	95.56	95.56	-	87.50	87.50	-	89.66	89.66	-
EntRel	58.73	-	58.73	70.97	-	70.97	47.06	-	47.06
Expansion.Alt	92.31	92.31	-	100.00	100.00	-	100.00	100.00	-
Expansion.Alt.Chosen alt	71.43	90.91	0.00	22.22	100.00	6.67	0.00	-	100.00
Expansion.Conjunction	70.45	97.00	40.00	75.88	98.36	40.26	63.48	94.52	27.51
Expansion.Instantiation	47.73	100.00	34.29	57.14	100.00	44.29	55.56	100.00	50.00
Expansion.Restatement	31.13	66.67	29.56	31.31	14.29	31.94	32.39	66.67	30.88
Temporal.Async.Precedence	78.46	98.00	13.33	82.22	100.00	11.11	84.44	97.44	0.00
Temporal.Async.Succession	82.83	87.23	0.00	58.82	63.49	0.00	96.08	96.08	-
Temporal.Synchrony	77.30	80.77	0.00	80.25	83.33	0.00	59.70	59.70	100.00
System				**All senses - comparison**					
Our system (Official)	64.13	91.20	40.32	**63.31**	89.80	**39.19**	54.69	78.34	34.56
Our improved system	64.77	91.05	41.66	62.69	90.02	37.81	**54.88**	78.38	34.92
Wang and Lan, 2015	**65.11**	90.00	**42.72**	61.27	**90.79**	34.45	54.76	76.44	36.29
Rutherford and Xue, 2016	-	-	40.32	-	-	36.13	-	-	**37.67**
Jain, 2016	62.43	**91.50**	36.85	50.90	89.70	15.60	41.47	**78.56**	9.95

Table 1: Evaluation of our official submission system, trained on Train 2016 and evaluated on Dev, Test and Blind sets. Comparison with our official system and our improved system with the official results of CoNLL 2015 Shared task's best system (Wang and Lan, 2015) and CoNLL 2016 Shared Task best systems in *Explicit* (Jain, 2016) and Non-Explicit (Rutherford and Xue, 2016). F-Score is presented.

system parameters as proposed in Kim (2014): filter windows with size 3,4,5 with 100 feature maps each, dropout probability 0.5 and mini-batch of size 50. We train the model with 50 epochs.

CNN ARC-1M experiments The *CNN ARC-1M* column shows results from our modification of ARC-1 CNN for sentence matching (see Section 3.3) fed with *Arg1* and *Arg2* word tokens' vector representations from the *Word2Vec* word embeddings. We use filter windows with size 3,4,5 with 100 feature maps each, shared between the two argument convolutions, dropout probability 0.5 and mini-batch of size 50 as proposed in Kim (2014). We train the model with 50 epochs.

Comparing *LR*, *CNN* and *CNN ARC-1M* according to their ability to classify different classes we observe that *CNN ARC-1M* performs best in detecting *Contingency.Cause.Reason* and *Contingency.Cause.Result* with a substantial margin over the other two models. The *CNN* model outperforms the *LR* and *CNN-ARC1M* for *Comparison.Contrast*, *EntRel*, *Expansion.Conjunction* and *Expansion.Instantiation* but cannot capture any *Expansion.Restatement* which leads to worse overall results compared to the others. These insights show that the Neural Network models are able to capture some dependencies between the relation arguments. For *Contingency.Cause.Results*, *CNN ARC-1M* even clearly outperforms the LR models enhanced with similarity features (discussed below). We also implemented a modified version of the *CNN ARC-2* architecture of Hu et al. (2015), which uses a cross-argument convolution layer, but it yielded much worse results.[7]

LR with Embeddings + Features The last three columns in Table 2 show the results of our feature-based Logistic Regression approach with different feature groups on top of the embedding representations of the arguments. Column *E+Sim* shows the results from our official submission and the other two columns show results for additional features that we added after the submission deadline.

Adding the cross-argument similarity features (without the POS modal verbs similarities) improves the overall result of the embeddings-only Logistic Regression (*LR*) baseline significantly from F-score 35.54 to 40.32. It also improves the result on almost all senses individually. Adding *Explicit connective similarities* features improves the *All* result by 0.67 points (E+Sim+Conn). It also improves the performance on *Tem-*

[7] We are currently checking our implementation.

	Embeddings only			Logistic Regression with Embeddings + Features		
Sense	**LR**	**CNN**	**CNN ARC-1M**	**E+Sim**	**E+Sim+Conn**	**E+Sim+Conn+POS MD**
Comparison.Concession	0.00	0.00	0.00	0.00	0.00	0.00
Comparison.Contrast	2.33	13.68	8.51	16.07	**18.80**	17.86
Contingency.Cause.Reason	25.00	29.30	35.90	38.51	40.24	**42.17**
Contingency.Cause.Result	3.57	9.20	**19.28**	15.38	15.38	13.70
EntRel	53.13	59.53	56.87	58.73	60.80	**61.26**
Expansion.Alt.Chosen alt	0.00	0.00	0.00	0.00	0.00	0.00
Expansion.Conjunction	35.90	38.29	14.67	40.00	40.91	**41.27**
Expansion.Instantiation	0.00	21.98	4.08	**34.29**	31.43	33.80
Expansion.Restatement	12.74	0.00	21.56	**29.56**	26.87	27.45
Temporal.Async.Precedence	0.00	0.00	0.00	13.33	**17.65**	12.90
Temporal.Async.Succession	0.00	0.00	0.00	0.00	0.00	0.00
Temporal.Synchrony	0.00	0.00	0.00	0.00	0.00	0.00
All	35.54	34.34	36.21	40.32	40.99	**41.66**

Table 2: Evaluation of different systems and feature configurations for Non-Explicit relation sense classification, trained on Train 2016 and evaluated on Dev. F-score is presented.

poral.Async.Precedence, *Expansion.Conjunction*, *EntRel*, *Contingency.Cause.Reason* and *Comparison.Contrast* individually. We further added *POS similarity features* between *MD (modal verbs)* and other part of speech tags between *Arg1* and *Arg2*. The obtained improvement of 0.67 points shows that the occurrence of modal verbs within arguments can be exploited for implicit discourse relation sense classification. Adding the modal verbs similarities also improved the individual results for the *Contingency.Cause.Reason*, *EntRel* and *Expansion.Conjunction* senses.

Some relations are hard to predict, probably due to the low distribution in the train and evaluation data sets: *Comparison.Concession*[8], *Expansion.Alt.Chosen alt*[9], *Temporal.Async. Succession*[10], *Temporal. Synchrony*[11].

5 Conclusion and Future work

In this paper we describe our system for the participation in the CoNLL Shared Task on Discourse Relation Sense Classification. We compare different approaches including Logistic Regression classifiers using features based on word embeddings and cross-argument similarity and two Convolutional Neural Network architectures. Our official submission uses a logistic regression classifier with several similarity features and performs with overall F-scores of 64.13 for the *Dev* set, 63.31 for the *Test* set and 54.69 for the *Blind* set. After the official submission we improved our system by adding more features for detecting senses for Non-Explicit relations and we improved our *Non-Explicit* result by 1.46 points to 41.66 on the *Dev* set and by 0.36 points to 34.92 on the *Blind* set.

We could show that dense representations of arguments and connectives jointly with cross-argument similarity features calculated over word embeddings yield competitive results, both for Explicit and Non-Explicit relations. First results in adapting CNN models to the task show that further gains can be obtained, beyond LR models.

In future work we want to explore further deep learning approaches and adapt them for discourse relation sense classification, using among others Recurrent Neural Networks and CNNs for matching sentences, as well as other neural network models that incorporate correlation between the input arguments, such as the MTE-NN system (Guzmán et al., 2016a; Guzmán et al., 2016b). Since we observe that the neural network approaches improve on the *LR* Embeddings-only models for most of the senses, in future work we could combine these models with our well-performing similarity features. Combining the output of a deep learning system with additional features has been shown to achieve state of the art performance in other tasks (Kreutzer et al., 2015).

Acknowledgments. This work is supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1.

We thank Ana Marasović for her advice in the implementation of CNN models.

[8] Comparison.Concession, Non-Explicit: Train:1.10 %, Dev:0.66 %: Test:0.59 %.

[9] Expansion.Alt.Chosen-alt, Non-Explicit: Train:0.79 %, Dev:0.26 %: Test:1.49 %.

[10] Temporal.Async.Succ, Non-Explicit: Train:0.80 %, Dev:0.39 %: Test:0.49 %.

[11] Temporal.Synchrony, Non-Explicit: Train:0.94 %, Dev:1.19 %: Test:0.49 %.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2201–2211, Lisbon, Portugal, September. Association for Computational Linguistics.

Christian Chiarcos and Niko Schenk. 2015. A minimalist approach to shallow discourse parsing and implicit relation recognition. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 42–49, Beijing, China, July. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, June.

Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2016a. Machine translation evaluation meets community question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, Berlin, Germany.

Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2016b. MTE-NN at SemEval-2016 Task 3: Can machine translation evaluation help community question answering? In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2015. Convolutional neural network architectures for matching natural language sentences. *CoRR*, abs/1503.03244.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

Fang Kong, Sheng Li, and Guodong Zhou. 2015. The sonlp-dp system in the conll-2015 shared task. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 32–36, Beijing, China, July. Association for Computational Linguistics.

Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. 2015. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 316–322, Lisbon, Portugal, September. Association for Computational Linguistics.

Sobha Lalitha Devi, Sindhuja Gopalan, Lakshmi S, Pattabhi RK Rao, Vijay Sundar Ram, and Malarkodi C.S. 2015. A hybrid discourse relation parser in conll 2015. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 50–55, Beijing, China, July. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland, June. Association for Computational Linguistics.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. *CoRR*, abs/1603.02776.

Ana Marasović and Anette Frank. 2016. Multilingual Modal Sense Classification using a Convolutional Neural Network. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, Berlin, Germany, August.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 368–375, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Todor Mihaylov and Preslav Nakov. 2016. SemanticZ at SemEval-2016 Task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 746–751, Atlanta, Georgia, USA.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alha Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In

Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, San Diego, California, USA.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90, Manchester, UK, August. Coling 2008 Organizing Committee.

Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2014. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Berlin Heidelberg New York, September. Springer.

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

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654, Gothenburg, Sweden, April. Association for Computational Linguistics.

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, Denver, Colorado, May–June. Association for Computational Linguistics.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Nat. Lang. Eng.*, 14(3):369–416, July.

Evgeny Stepanov, Giuseppe Riccardi, and Ali Orkan Bayer. 2015. The unitn discourse parser in conll 2015 shared task: Token-level sequence labeling with argument-specific models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 25–31, Beijing, China, July. Association for Computational Linguistics.

Quan Hung Tran, Vu Tran, Tu Vu, Minh Nguyen, and Son Bao Pham. 2015. JAIST: Combining multiple features for answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 215–219, Denver, Colorado, USA.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24, Beijing, China, July. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China, July. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Bonnie Webber, Attapol Rutherford, Chuan Wang, and Hongmin Wang. 2016. The conll-2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

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 the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Lisbon, Portugal, September. Association for Computational Linguistics.