Zero-shot transfer for implicit discourse relation classification

Murathan Kurfalı
Department of Linguistics
Stockholm University
murathan.kurfali@ling.su.se

Robert Östling
Department of Linguistics
Stockholm University
robert@ling.su.se

Abstract

Automatically classifying the relation between sentences in a discourse is a challenging task, in particular when there is no overt expression of the relation. It becomes even more challenging by the fact that annotated training data exists only for a small number of languages, such as English and Chinese. We present a new system using zero-shot transfer learning for implicit discourse relation classification, where the only resource used for the target language is unannotated parallel text. This system is evaluated on the discourse-annotated TED-MDB parallel corpus, where it obtains good results for all seven languages using only English training data.

1 Introduction

The difference between a set of randomly selected sentences and a discourse lies in coherence. Among other attempts at defining the elusive nature of coherence, one way is to look at the meaning conveyed between the adjacent pair of sentences. In the current study, we follow the Penn Discourse Treebank (PDTB) framework which regards abstract objects (Asher, 2012) as the units of discourse and views the text as a collection of discourse level predicates, each taking two arguments. Such predicates, called discourse connectives, may (Ex. 1) or may not (Ex. 2) be represented in the surface form:

1. <u>Because</u> **the drought reduced U.S. stockpiles**, *they have more than enough storage space for their new crop*, and that permits them to wait for prices to rise.

2. *But a few funds have taken other defensive steps. Some have raised their cash positions to record levels.* Implicit = BECAUSE **High cash positions help buffer a fund when the market falls**.

where *italics* represents the first and **boldface** the second argument to the <u>underlined</u> discourse connective. The discourse relations which lack an overt discourse connective (Ex. 2) are referred as *implicit discourse relations* and are shown to be the most challenging part of the discourse parsing (e.g. Pitler et al., 2009).

In this paper, we perform implicit discourse relation classification using three recent data sets annotated according to the same guidelines: Penn Discourse Treebank (PDTB) 3.0, the Turkish Discourse Bank (TDB), and the multilingual TED-MDB. To the best of our knowledge, multilingual training and zero-shot transfer has not previously been investigated for this problem. The results suggest that an implicit discourse relation classifier can transfer well across dissimilar languages, and that pooling training data from unrelated languages (English and Turkish) leads to significantly better performance for all languages.

2 Related Work

Implicit discourse relation recognition is often handled as a classification task, where earlier studies focused on using linguistically rich features (Pitler et al., 2009; Zhou et al., 2010; Park and Cardie, 2012; Rutherford and Xue, 2014).

Recently, neural network approaches have become popular. Ji and Eisenstein (2015) use two RNNs on the syntactic trees of the arguments whereas Zhang et al. (2015) use a CNN to perform discourse parsing in a multi-task setting where they consider both explicit and implicit discourse relations.

Rutherford and Xue (2016) use a simple yet robust feedforward network and achieves the highest performance on the out-of-domain blind test in the CoNLL 2016 shared task (Xue et al., 2016).

Lan et al. (2017) apply a multi-task attention-

226

based neural network model whereas Bai and Zhao (2018) focus on the representation of the sentence pair and take different levels of text, from character to sentence pair, into account to achieve a richer representation.

Dai and Huang (2018) adopt a similiar approach and represent discourse units by considering a wider paragraph-level context. The discourse unit representations are created by a Bi-LSTM which takes a sequence of discourse relations in a paragraph which enables capturing the inter-dependencies between discourse relations as well.

3 Data

We use four different data sets: the Penn Discourse Treebank (PDTB) version 2.0 (Prasad et al., 2008) and version 3.0 (Prasad et al., 2018), as well as the TED Multilingual Discourse Bank (TED-MDB, Zeyrek et al. 2018) and the Turkish Discourse Bank (TDB, Zeyrek and Kurfalı 2017).

The PDTB is built upon the 1 million word Wall Street Journal corpus and is the largest available resource for discourse relations. Most related work uses PDTB 2.0, so we include this for comparing our baseline to previous work.

The recently released PDTB 3.0 adopts a new annotation schema as well as an updated sense hierarchy. PDTB 3.0 includes the annotations of PDTB 2.0 updated according to the new annotation schema, as well as about 13 thousand new annotations, of which about 5K are implicit relations (Prasad et al., 2018). The distribution of the top level senses of the implicit discourse relations in both PDTB versions is provided in Table 1.

TED-MDB (Zeyrek et al., 2018) is the first parallel corpus annotated for discourse relations. It closely follows the PDTB 3.0 framework and includes the manual annotations of six TED talks in seven languages (English, Turkish, European Portuguese, Polish, German, Russian) aiming to allow crosslingual comparison of discourse relations[1]. It has recently also been updated with Lithuanian (Oleskeviciene et al., 2019).

Despite the high number of languages covered by TED-MDB, the amount of annotated text per language is limited (see Table 2). Therefore, in the current study, we limit ourselves with the top level senses, namely Expansion, Contingency, Compar-

ison and Temporal. We only use TED-MDB for evaluation.

Among the TED-MDB languages other than English, only Turkish has another corpus annotated with PDTB 3.0 discourse annotations, namely the Turkish Discourse Bank (TDB). TDB is a multi-genre corpus of 40 000 words, considerably less than the PDTB (see Table 2), but it provides the only directly comparable baseline to assess the performance of zero-shot learning.

Sense	PDTB2			PDTB3		
	Train	Dev	Test	Train	Dev	Test
Comp.	1894	401	146	1828	404	153
Cont.	3281	628	276	5872	1159	527
Exp.	6792	1253	556	7939	1466	643
Temp.	665	93	68	1413	230	148

Table 1: Distribution of top level senses of the implicit discourse relations in PDTB 2.0 and PDTB 3.0 training, development and test sets: comp(arison), cont(ingency), exp(ansion), temp(oral).

4 Model

The main purpose of this study is to assess the performance of transfer learning on the implicit discourse relation classification task. To this end, we use a simple feedforward network fed with multilingual sentence embeddings following the finding of (Rutherford et al., 2017) which shows that simple discourse models with feedforward layers perform on par or better than those of with surface features or recurrent and convolutional architectures.

We follow the model of (Rutherford and Xue, 2016) due to its simplicity and robust nature even in the multilingual setting with different argument and discourse relation representations. We represent the arguments of the discourse relation via pre-trained LASER model (Artetxe and Schwenk, 2018). LASER is chosen as it is the current state-of-the-art model on several Natural Language Inference (NLI) transfer learning tasks, a sentence relation classification problem similar to discourse relation classification.

Given the argument vectors, V_{arg1} and V_{arg2}, the next step is to represent the discourse relation in a way that the interactions between them are captured. To this end, we model the discourse relation vector, V_{dr}, by performing the following pair-wise vector operations following the DisSent model of (Nie et al., 2017):

---

[1]The TED-MDB annotations are available at: https://github.com/MurathanKurfali/Ted-MDB-Annotations

Language	Comparison	Contingency	Expansion	Temporal	Total
English	20 (10.31%)	52 (26.80%)	107 (55.15%)	15 (7.73%)	194 (100%)
German	13 (6.07%)	41 (19.16%)	148 (69.16%)	12 (5.61%)	214 (100%)
Lithuanian	26 (10.57%)	53 (21.54%)	154 (62.60%)	13 (5.28%)	246 (100%)
Polish	19 (9.74%)	28 (14.36%)	130 (66.67%)	18 (9.23%)	195 (100%)
Portuguese	23 (9.06%)	47 (18.50%)	169 (66.54%)	15 (5.91%)	254 (100%)
Russian	16 (7.24%)	31 (14.03%)	169 (76.47%)	5 (2.26%)	221 (100%)
Turkish	20 (9.90%)	29 (14.36%)	140 (69.31%)	13 (6.44%)	202 (100%)
TDB (training)	71 (10.94%)	142 (21.88%)	363 (55.93%)	73 (11.25%)	649 (100%)
TDB (dev)	11 (9.82%)	31 (27.68%)	49 (43.75%)	21 (18.75%)	112 (100%)

Table 2: Distribution of top level senses of the implicit discourse relations in the TED-MDB and TDB corpora. The numbers within the parenthesis indicate the ratio. Since there is no official training/dev split for TDB, we arbitrarily chose two sections with different genres for the development set.

$$V_{avg} = \frac{1}{2}(V_{arg1} + V_{arg2})$$

$$V_{sub} = V_{arg1} - V_{arg2}$$

$$V_{mul} = V_{arg1} * V_{arg2}$$

$$V_{dr} = [V_{arg1}, V_{arg2}, V_{avg}, V_{sub}, V_{mul}]$$

The resulting vector is further fed into a hidden layer h_t with d hidden units[2] to achieve a more abstract representation of the relation and finally the output o is calculated using the sigmoid function. This model is also essentially the same as was used by Artetxe and Schwenk (2018) for NLI transfer learning.

5 Experiments

We formulate the implicit relation classification as four "one vs other" binary classification task. We follow the conventional setting of the first study (Pitler et al., 2009) and split the PDTB 2.0 into training (sections 2-20), development (sections 0-1 and 23-24) and test sets (sections 21-22) to have directly comparable results with the previous work. However, following the PDTB's original distinction but unlike some previous work, we distinguish Entity-based relations from implicit relations. Each classifier is trained on an equal number of positive and negative instances where the negative instances are randomly selected in each epoch to have a better representation of the data during the training. This model is evaluated on the PDTB 2.0 test set to confirm whether our model performs adequately on same-language, same-domain data. These results are directly comparable to previous work.

As TED-MDB is annotated according to the PDTB 3.0 framework, we train separate classifiers on PDTB 3.0 following the same convention as above. We test the trained models on all the implicit discourse relations in the TED-MDB corpus.

The PDTB framework allows annotations to be labelled with more than one label. In such cases we only keep the first label, in line with previous studies (among others Ji and Eisenstein, 2015; Rutherford et al., 2017).

The argument vectors are kept fixed during the training, and we do not update the parameters of the LASER model. We use cross-entropy loss, and AdaGrad as the optimizer. We evaluate using the model which achieved the highest F-score on the development set. As for the regularization, we use a dropout layer between the input and the hidden layer with a dropout probability of 0.3. All models are run 100 times to estimate the variance due to random initialization and stochastic training. All the models are implemented in PyTorch[3].

6 Results and Discussion

Table 3 shows the same-language, same-domain performance of our system, in comparison to previous work. All figures refer to PDTB 2.0 test set F-score, when trained on the PDBT 2.0 training set, and are directly comparable. While our model does not achieve state-of-the-art performance in this setting, this experiment shows that it performs adequately for English, and provides a reasonable baseline for the zero-shot experiments presented in Tables 4 and 5. We also include a naive baseline system which always predicts TRUE and is evaluated on the respective (PDTB 2.0 or PDTB 3.0)

---

[2] We use d=100 in the experiments

[3] https://pytorch.org/

	Comparison	Contingency	Expansion	Temporal
(Pitler et al., 2009)	21.96	47.13	-	16.76
(Zhou et al., 2010)	31.79	47.16	70.11	20.30
(Park and Cardie, 2012)	31.32	49.82	-	26.57
(Rutherford and Xue, 2014)	39.70	54.42	70.23	28.69
(Zhang et al., 2015)	33.22	52.04	69.59	30.54
(Ji and Eisenstein, 2015)	35.93	52.78	-	27.63
(Lan et al., 2017)	40.73	**58.96**	**72.47**	38.50
(Bai and Zhao, 2018)	**47.85**	54.47	70.60	36.87
(Dai and Huang, 2018)	46.79	57.09	70.41	**45.61**
Baseline	24.49	41.75	69.41	12.20
Our system	28.19 (±0.83)	50.63 (±1.00)	64.07 (±1.90)	29.22 (±2.53)

Table 3: Comparison of the F scores (%) of binary classifiers on PDTB 2.0 test set. Left out scores refer to the results where EntRel relations are also considered to be Expansion.

Language	Comparison	Contingency	Expansion	Temporal	Average
Baseline (PDTB 3.0)	18.84	52.75	60.83	18.28	37.67
PDTB 3.0	24.90 (±0.87)	59.18 (±0.72)	60.10 (±1.32)	36.73 (±1.45)	45.23
German	8.62 (±1.61)	37.34 (±1.43)	70.81 (±3.16)	40.11 (±4.32)	39.22
English	10.18 (±3.31)	40.92 (±1.80)	62.28 (±2.16)	50.45 (±5.26)	40.96
Lithuanian	23.50 (±2.33)	34.64 (±1.43)	62.35 (±2.65)	36.78 (±3.28)	39.32
Polish	16.50 (±3.51)	29.19 (±1.36)	60.32 (±2.84)	44.17 (±3.37)	37.54
Portuguese	19.59 (±1.99)	33.85 (±1.27)	66.83 (±2.57)	37.04 (±3.43)	39.33
Russian	14.90 (±2.07)	26.76 (±1.08)	70.06 (±3.97)	28.28 (±4.41)	35.00
Turkish	10.99 (±3.16)	25.28 (±1.23)	64.14 (±2.96)	33.66 (±4.31)	33.52

Table 4: F scores (%) when the model is trained only on PDTB 3.0. In the table, PDTB 3.0 refers to the test set of the PDTB 3.0 corpus. The remaining rows refer to evaluations using TED-MDB.

test set in our comparisons.

In all zero-shot experiments, evaluation is performed on the available test data with PDTB 3.0 annotations: TED-MDB, and the PDTB 3.0 test set itself. Results in Table 4 use PDTB 3.0 only for training, whereas Table 5 presents the effect of having additional training data from Turkish (a language unrelated to English). Pooling training data from different languages is possible since our model is language-agnostic.

In all zero-shot experiments, we see similar levels of performance across all the evaluated languages in TED-MDB. While not completely comparable numerically since annotations differ slightly between languages, this evaluation set consists of parallel sentences annotated according to the same guidelines. The similarity in scores between the training language(s)—English and/or Turkish—and the remaining languages indicates that little accuracy is lost during transfer.

Comparing the performances with and without additional Turkish data, TDB, reveals that adding a small amount (relative to the size of PDTB 3.0) of Turkish training data improves the F-scores by a statistically significant amount[4] for not only Turkish, but for all the languages in TED-MDB Table 5.

7 Conclusion

In the current paper we have presented the (to the best of our knowledge) first study of zero-shot learning in the implicit discourse relation classification task. Our method does not require any discourse level annotation for the target languages, yet still achieves good performance even for those languages where no training data is available. The performance is further increased by pooling training data from multiple languages. Using our published code[5] and publicly available resources it can used for implicit discourse classification in

---

[4]On a sense-wise analysis, we observe that the main increase is in the Expansion relations; however, there is no decrease in any of the other senses.

[5]https://github.com/MurathanKurfali/multilingual_IDRC

Language	TDB	PDTB3	PTDB3+TDB
PDTB3 Test	35.35	45.23	**45.62**
German	36.93	39.22	**41.44**
English	38.06	40.96	**42.22**
Lithuanian	36.92	39.32	**41.94**
Polish	35.48	37.54	**39.65**
Portuguese	37.58	39.33	**41.04**
Russian	30.92	35.00	**38.23**
Turkish	**39.58**	33.52	37.14

Table 5: Comparison of average F-scores (%) when the model is trained on different training sets. Bold means significantly higher F-score than the second highest column ($p < 0.001$, Mann-Whitney U test).

nearly a hundred languages.

Acknowledgments

We would like to thank Bonnie Webber for her help in obtaining PDTB 3.0 and Mats Wirén for his useful comments.

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *arXiv preprint arXiv:1812.10464*.

Nicholas Asher. 2012. *Reference to abstract objects in discourse*, volume 50. Springer Science & Business Media.

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. *arXiv preprint arXiv:1807.05154*.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. *arXiv preprint arXiv:1804.05918*.

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, 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.

Allen Nie, Erin D Bennett, and Noah D Goodman. 2017. Dissent: Sentence representation learning from explicit discourse relations. *arXiv preprint arXiv:1710.04334*.

Giedre Valunaite Oleskeviciene, Deniz Zeyrek, Viktorija Mazeikiene, and Murathan Kurfalı. 2019. Observations on the annotation of discourse relational devices in ted talk transcripts in lithuanian. *Proceedings of the workshop on annotation in digital humanities co-located with ESSLLI 2018*, pages 53–58.

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.

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.

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, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the pdtb: The next generation. In *Proceedings 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291.

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.

Attapol Rutherford and Nianwen Xue. 2016. Robust non-explicit neural discourse parser in english and chinese. *Proceedings of the CoNLL-16 shared task*, pages 55–59.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. Conll 2016 shared task on multilingual shallow discourse parsing. *Proceedings of the CoNLL-16 shared task*, pages 1–19.

Deniz Zeyrek and Murathan Kurfalı. 2017. Tdb 1.1: Extensions on turkish discourse bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2018. Ted multilingual discourse bank (ted-mdb): a parallel corpus annotated in the pdtb style. *Language Resources and Evaluation*, pages 1–27.

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.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1507–1514. Association for Computational Linguistics.