# A Unified Approach to Discourse Relation Classification in nine Languages

**Hanna Varachkina**
University of Göttingen, Germany
Department of German Philology
hanna.varachkina@
stud.uni-goettingen.de

**Franziska Pannach**
University of Göttingen, Germany
Göttingen Centre for Digital Humanities
franziska.pannach@
uni-goettingen.de

## Abstract

This paper presents efforts to solve the shared task on discourse relation classification (disrpt task 3). The intricate prediction task aims to predict a large number of classes from the Rhetorical Structure Theory (RST) framework for nine target languages. Labels include discourse relations such as *background*, *condition*, *contrast* and *elaboration*. We present an approach using euclidean distance between sentence embeddings that were extracted using multlingual sentence BERT (sBERT) and directionality as features. The data was combined into five classes which were used for initial prediction. The second classification step predicts the target classes. We observe a substantial difference in results depending on the number of occurrences of the target label in the training data. We achieve the best results on Chinese, where our system achieves 70 % accuracy on 20 labels.

## 1 Introduction

Discourse relations are an integral part of natural language understanding. They provide information on the interaction between aspects of utterances, such as *result* “I had already seen the movie, so the ending of the book was spoiled!” or *contrast* “I did not like her first book, but her latest novel is great!”. In natural language processing, understanding discourse relations is beneficial for many tasks, e.g. the classification of sentiment (Benamara et al., 2016).

Due to different frameworks and the complexity of the subject, the classification of discourse relations is a challenging task that has been addressed by the Shared Task on Discourse Relation Classification across Formalisms (Task 3 of the workshop on Discourse Relation Parsing and Treebanking (DISRPT))[1]. The proposed shared task is to our knowledge the first of its kind on discourse relation classification across formalisms.

In the past, a number of classification efforts for the data created within the PDTB framework and RST (Hernault et al., 2010; Liu et al., 2016; Kim et al., 2020), mostly for English have been attempted. This paper focuses on the data based on the Rhetorical Structure Theory (RST) of nine languages: Basque, Chinese, Dutch, English, Farsi, German, Portuguese, Russian and Spanish in a cross-lingual approach.

## 2 Data

The data sets created on the basis of RST were used for the task of discourse relation classification, since the text data for most other data sets was hidden by underscores and we did not have resources to restore it. Out of two data sets provided for Spanish the RST Spanish Treebank (spa.rst.rststb) was chosen. Due to the lack of the original text data, the rows with units hidden by underscores were removed from the English data set (Georgetown University Multilayer corpus – eng.rst.gum). This reduced English model was trained for comparison with other models.

The RST corpora were analyzed and compared. We found that they have different size, e.g. the Chinese data set is the smallest and the Russian data set exceeds the size of all the other data sets.

The corpora for different languages differ in their number of labels. Only the labels *background*, *condition*, *contrast* and *elaboration* appear in all languages. Table 1 shows the distribution of labels between languages based on their training, development and test sets combined. The second to last column shows unique labels, 57 in total for all languages. Some labels of different languages differ slightly in spelling. Those were considered as one (e.g. *antithesis* and *anthitesis*). In Span-

[1] https://sites.google.com/georgetown.edu/disrpt2021

*Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 46–50
November 11, 2021. ©2021 Association for Computational Linguistics

| German | English | Basque | Farsi | Dutch | Portuguese | Russian | Spanish | Chinese | Unique labels | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | alternative | | alternative | 0 |
| antithesis | antithesis | antithesis | | antithesis | antithesis | antithesis | antithesis | antithesis | antithesis | 0 |
| | attribution | | attribution | | attribution | attribution | | attribution | attribution | 4 |
| background | background | background | background | background | background | background | backgroun(d) | background | background | 4 |
| cause | cause | cause | cause | | | cause | cause | cause | cause | 2 |
| | | | | | | cause-effect | | | cause-effect | 2 |
| circumstance | circumstance | circumstance | | circumstance | circumstance | | circumstance | circumstance | circumstance | 4 |
| | | | comparison | | comparison | comparison | | | comparison | 0 |
| concession | concession | concession | | concession | concession | concession | concession | concession | concession | 0 |
| | | | | | conclusion | conclusion | | | conclusion | 2 |
| condition | condition | condition | condition | condition | condition | condition | condition | condition | condition | 0 |
| conjunction | | conjunction | | conjunction | | | conjunction | conjunction | conjunction | 3 |
| contrast | contrast | contrast | contrast | contrast | contrast | contrast | contrast | contrast | contrast | 0 |
| disjunction | | disjunction | | disjunction | | | disjunction | disjunction | disjunction | 3 |
| | | | | | | effect | | | effect | 2 |
| e-elaboration | | | | | | | | | e-elaboration | 4 |
| elaboration | elaboration | elaboration | elaboration | elaboration | elaboration | elaboration | elaboration | elaboration | elaboration | 4 |
| | | enablement | enablement | enablement | enablement | | enablement | enablement | enablement | 4 |
| | evaluation | evaluation | evaluation | evaluation | evaluation | evaluation | evaluation | evaluation | evaluation | 1 |
| evaluation-n | | | | | | | | | evaluation-n | 1 |
| evaluation-s | | | | | | | | | evaluation-s | 1 |
| evidence | evidence | evidence | | evidence | evidence | evidence | evidence | evidence | evidence | 4 |
| | | | explanation | | explanation | | | | explanation | 4 |
| interpretation | | interpretation | | interpretation | interpretation | | interpretation | interpretation | interpretation | 1 |
| | | | | | | interpretation-evaluation | | | interpretation-evaluation | 1 |
| joint | joint | joint | joint | joint | joint | joint | joint | | joint | 3 |
| | justify | justify | | justify | justify | | justify | justify | justify | 1 |
| list | | list | | list | list | | list | list | list | 3 |
| | manner | | | | | | | | manner | 2 |
| | | | manner-means | | | | | | manner-means | 2 |
| means | means | means | | means | means | | means | means | means | 2 |
| | motivation | motibation | | motivation | motivation | motivation | motivation | motivation | motivation | 2 |
| | | | | nonvolitional-cause | non-volitional-cause | | | | non-volitional-cause | 2 |
| | | | | | non-volitional-cause-e | | | | non-volitional-cause-e | 2 |
| | | | | nonvolitional-result | non-volitional-result | | | | non-volitional-result | 2 |
| | | | | | non-volitional-result-e | | | | non-volitional-result-e | 2 |
| | | otherwise | | otherwise | otherwise | | | | otherwise | 0 |
| | | | | | parenthetical | | | | parenthetical | 4 |
| preparation | preparation | preparation | | preparation | | preparation | preparation | preparation | preparation | 4 |
| purpose | purpose | purpose | | purpose | purpose | purpose | purpose | purpose | purpose | 2 |
| | question | | | | | | | | question | 4 |
| reason | | | | | | | | | reason | 2 |
| restatement | restatement | restatement | | restatement | restatement | restatement | restatement | restatement | restatement | 4 |
| | | | | restatement-mn | | | | | restatement-mn | 4 |
| result | result | result | | | | | result | result | result | 2 |
| sequence | sequence | sequence | | sequence | sequence | sequence | sequence | sequence | sequence | 3 |
| solutionhood | solutionhood | solution-hood | | solutionhood | solutionhood | solutionhood | solutionhood | solutionhood | solutionhood | 4 |
| | | | | span | | | | | span | 3 |
| summary | | summary | summary | summary | | | summary | summary | summary | 4 |
| | | | temporal | | | | | | temporal | 3 |
| | | | topichange | | | | | | topichange | 4 |
| | | | topicomment | | | | | | topicomment | 4 |
| | | | topidrift | | | | | | topidrift | 4 |
| | | unconditional | | unconditional | | | | | unconditional | 0 |
| | | unless | | unless | | | unless | | unless | 0 |
| | | | | volitional-cause | volitional-cause | | | | volitional-cause | 2 |
| | | | | volitional-result | volitional-result | | | | volitional-result | 2 |

Table 1: All labels and their assignment to five large categories: 0, 1, 2, 3, 4.

ish, *backgroun* in the train and *background* in the development sets were considered as one label.

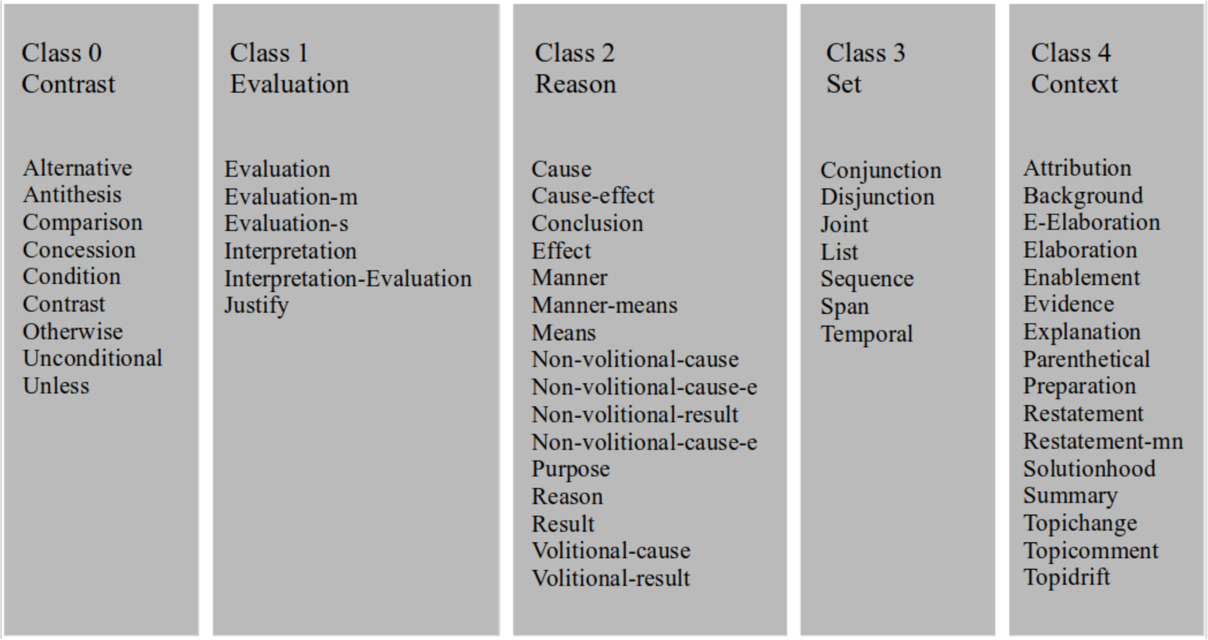
In all cases except English (after removing rows with underscores), the number of labels in the test set is smaller than in the train and development sets combined (see Table 2). For Portuguese, this difference is the largest with the test set containing 11 labels less than the training set. In case of Dutch, the test data contains one instance with the label *span* that is not present neither in the train nor in the development set. Apart from that, all labels in the test sets of the other languages occur in their train and/or the development sets and can be learnt.

The number of labels in each data set is large (between 16 and 29), so that it is quite difficult for any classifier to distinguish between them. Inspired by other discourse relation schemes such as PDTB (Prasad et al., 2008) that suggests hierarchical labels, unique labels were mapped to five large classes (last column in Table 1) as some of them can be seen as subclasses of the other. This was done in order to break down the task of classifying a very large number of classes as will be described in the next section. Figure 1 shows which classes are grouped together.

## 3 Method

The subject of discourse relation classification are not text units, but rather relations between them. One of the methods for comparing two text units is to calculate euclidean distance between their embeddings. Using multilingual sBERT (Reimers and Gurevych, 2020) models, embeddings for two discourse units were obtained [2]. The first version of the model created for 15 languages (Yang et al., 2020) was used for the corresponding languages in the shared task data set. Its second version for a larger number of languages was used for the re-

[2] https://www.sbert.net/docs/pretrained_models.html



| Class 0<br>Contrast | Class 1<br>Evaluation | Class 2<br>Reason | Class 3<br>Set | Class 4<br>Context |
|---|---|---|---|---|
| Alternative | Evaluation | Cause | Conjunction | Attribution |
| Antithesis | Evaluation-m | Cause-effect | Disjunction | Background |
| Comparison | Evaluation-s | Conclusion | Joint | E-Elaboration |
| Concession | Interpretation | Effect | List | Elaboration |
| Condition | Interpretation-Evaluation | Manner | Sequence | Enablement |
| Contrast | Justify | Manner-means | Span | Evidence |
| Otherwise | | Means | Temporal | Explanation |
| Unconditional | | Non-volitional-cause | | Parenthetical |
| Unless | | Non-volitional-cause-e | | Preparation |
| | | Non-volitional-result | | Restatement |
| | | Non-volitional-cause-e | | Restatement-mn |
| | | Purpose | | Solutionhood |
| | | Reason | | Summary |
| | | Result | | Topichange |
| | | Volitional-cause | | Topicomment |
| | | Volitional-result | | Topidrift |

Figure 1: Five classes used for first level training

maining languages in the data sets, since the second model works slightly worse on the languages from the first model. Then euclidean distance between the embeddings of these two units was calculated. The euclidean distance is the first feature for the model and the second is a categorical one taken from the column that indicate the directionality information. Initially, we experimented with several embeddings for the German corpus, but the results were similar, including those for the multilingual embeddings. Therefore, we used the same multilingual model for several languages for simplicity reasons.

Using the features mentioned above, a stacked random forest model was trained with one-versus-all approach and balanced weights for classes. The model parameters were the same for all the layers of the stacked model. Since the number of original classes is very large, the labels were grouped into larger classes as the last column in Table 1 shows. The original training and development sets were combined and split, so that the development set comprises 40 % of the combined one. This development set was used for testing a random forest model for five upper classes. After training the model, the predictions for large classes were mapped to the original labels. Now the test set of the classifier for large classes (the development set after resizing) was used for training five models, one for each large class that classify the original labels contained in the large class. The predictions of five models were compared with the original test data for evaluation. As for the Chinese data set, the original training set was duplicated into the development set in order to avoid spareness, while all the other steps in the model were the same as described above.

## 4 Results

Poorly predicted large classes set restrictions on further label classification. Some of the classes have very few instances so that models cannot learn and recognize them properly. Table 4 shows F1-score results for individual labels in the Chinese data set. Out of 20 original classes in the test data, 10 were not recognized correctly, i.e. exactly a half, which is one of the highest numbers of neglected classes among languages. This can be explained by the fact that the number of train instances for these classes is extremely low. Russian (Table 5) is the language with the smallest number of labels that were not recognized (only 2). It is also the largest data set with a sufficient number of instances for most labels to train and test a model on.

In general, the smaller the number of classes and the larger the data set (and the number of instances for each label), the better is the accuracy score. As can be seen in Table 2, the results range from 35 % accuracy for German to 70 % for Chinese. Although half of the labels for Chinese were not recognized, the remaining labels were recognized very well, which led to a high overall accuracy. For example, the *elaboration* label has 69 instances in the test set and receives 93 % F1-score, while the *evidence* label has only one instance and receives 0 % F1-score, see Table 4.

| Language | Test set length | Unique labels | Test labels | Acc. |
|---|---|---|---|---|
| German | 260 | 26 | 24 | 0.35 |
| English | 1800 | 23 | 23 | 0.49 |
| Basque | 667 | 29 | 26 | 0.44 |
| Farsi | 592 | 17 | 16 | 0.59 |
| Dutch | 324 | 33 | 29 | 0.45 |
| Portuguese | 263 | 32 | 21 | 0.49 |
| Russian | 2838 | 22 | 20 | 0.61 |
| Spanish | 426 | 28 | 25 | 0.45 |
| Chinese | 159 | 26 | 20 | 0.70 |

Table 2: Evaluation results on RST data sets. The English data is taken after removing underscores.

The organizers of the shared task evaluated our model on all data sets, including non-RST data that we did not have access to due to accessibility restrictions. Additional labels from the non-RST data sets should be regarded separately from the labels in Table 1. Since we did not have text data, we decided not to judge whether they fit into the thematic categories in Figure 1. The organizers assigned these additional labels successively to each category (so the first additional relation gets group 1, the next gets group 2, and so on in a rotation). The results can be observed in Table 3. The data sets with anonymized data were included into evaluation and the English data GUM that was partially anonymized was included entirely. Beside the sparse Chinese data set (zho.rst.sctb), the Spanish version of the same corpus (spa.rst.sctb) as well as the French data set (fra.sdrt.annodis) are also sparse. Therefore, the training set was duplicated into the development set. As for the results, the accuracy for the English data set including the hidden text units drops from 49 % to 47 %. As for the Spanish data sets, the score for the corpus that we chose (spa.rst.rststb) is lower than for the second Spanish corpus (spa.rst.sctb), 45 % compared to 69 %. On average, we reach 54 % accuracy.

| Test set | Acc. |
|---|---|
| **deu.rst.pcc** | 0.35 |
| eng.pdtb.pdtb | 0.50 |
| **eng.rst.gum** | 0.47 |
| eng.rst.rstdt | 0.55 |
| eng.sdrt.stac | 0.54 |
| **eus.rst.ert** | 0.44 |
| **fas.rst.prstc** | 0.59 |
| fra.sdrt.annodis | 0.46 |
| **nld.rst.nldt** | 0.45 |
| **por.rst.cstn** | 0.49 |
| **rus.rst.rrt** | 0.61 |
| **spa.rst.rststb** | 0.45 |
| spa.rst.sctb | 0.69 |
| tur.pdtb.tdb | 0.48 |
| zho.pdtb.cdtb | 0.89 |
| **zho.rst.sctb** | 0.70 |
| Average | 0.54 |

Table 3: Evaluation results from the organizers. The RST data sets from Table 2 are in bold.

## 5 Conclusion

Although the data taken for this shared task is based on the Rhetorical Structure Theory, each language (and corpus) proposes a different set of labels for classification. However, in many cases there are substantial overlaps which makes a unified approach to discourse relation classification possible.

This paper suggests such an approach by grouping all unique labels, some of which are shared between languages, to large classes and using a stacked random forest model in each case. However, a different division into larger groups, for example using automatic approaches such as clustering could be beneficial and reduce the amount of manual work. Further fine-tuning of the embeddings with the shared task data will potentially improve the results. As for the random forest method, different hyperparameters for different languages can be tried out in additional experiments.

We show that our approach achieves good results on labels with a sufficient amount of instances per class across languages. Albeit the simplicity of the model and small number features, we achieve 70 % accuracy on Chinese labels and at least 35 % for German.

| Label<br>(test) | F1-score<br>(test) | Support<br>(test) | Support<br>(dev set) |
|---|---|---|---|
| antithesis | 0.40 | 3 | 5 |
| background | 0.00 | 4 | 13 |
| circumstance | 0.00 | 4 | 4 |
| condition | 1.00 | 1 | 6 |
| conjunction | 0.00 | 2 | 3 |
| contrast | 0.33 | 5 | 2 |
| disjunction | 0.00 | 2 | 4 |
| elaboration | 0.93 | 69 | 114 |
| enablement | 0.00 | 1 | 1 |
| evidence | 0.00 | 1 | 3 |
| interpretation | 0.50 | 3 | 5 |
| list | 0.84 | 32 | 80 |
| means | 0.00 | 2 | 7 |
| motivation | 0.00 | 1 | 2 |
| preparation | 0.73 | 12 | 48 |
| purpose | 0.36 | 6 | 11 |
| restatement | 0.00 | 1 | 0 |
| result | 0.25 | 4 | 8 |
| sequence | 0.22 | 5 | 12 |
| summary | 0.00 | 1 | 5 |
| additional label in dev set | | | |
| cause | 0.00 | 0 | 6 |
| justify | 0.00 | 0 | 3 |
| concession | 0.00 | 0 | 8 |
| solutionhood | 0.00 | 0 | 1 |
| | | 159 | 351 |

Table 4: Classification results for Chinese

| Label<br>(test) | F1-score<br>(test) | Support<br>(test) | Support<br>(dev set) |
|---|---|---|---|
| antithesis | 0.00 | 2 | 12 |
| attribution | 0.18 | 58 | 688 |
| background | 0.12 | 68 | 299 |
| cause | 0.58 | 208 | 930 |
| cause-effect | 0.04 | 27 | 123 |
| comparison | 0.08 | 41 | 236 |
| concession | 0.06 | 27 | 156 |
| condition | 0.59 | 173 | 635 |
| contrast | 0.61 | 202 | 850 |
| effect | 0.00 | 1 | 27 |
| elaboration | 0.82 | 701 | 2899 |
| evaluation | 0.86 | 135 | 520 |
| evidence | 0.04 | 73 | 329 |
| interpretation<br>-evaluation | 0.14 | 15 | 65 |
| joint | 0.83 | 671 | 3017 |
| preparation | 0.49 | 149 | 671 |
| purpose | 0.31 | 92 | 506 |
| restatement | 0.08 | 24 | 114 |
| sequence | 0.16 | 150 | 455 |
| solutionhood | 0.08 | 26 | 156 |
| additional label in dev set | | | |
| conclusion | 0.00 | 0 | 2 |
| | | 2843 | 12690 |

Table 5: Classification results for Russian

## References

Farah Benamara, Maite Taboada, and Yannick Mathieu. 2016. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43:1–103.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. Towards semi-supervised classification of discourse relations using feature correlations. In *Proceedings of the SIGDIAL 2010 Conference*, pages 55–58.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

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 *Proceedings of the 6th Language Resources and Evaluation Conference*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.