Acquiring Annotated Data with Cross-lingual Explicitation for Implicit Discourse Relation Classification

Wei Shi[†], **Frances Yung**[†] and **Vera Demberg**[†,‡]

[†]Dept. of Language Science and Technology
[‡]Dept. of Mathematics and Computer Science, Saarland University
Saarland Informatic Campus, 66123 Saarbrücken, Germany
{w.shi, frances, vera}@coli.uni-saarland.de

Abstract

Implicit discourse relation classification is one of the most challenging and important tasks in discourse parsing, due to the lack of connectives as strong linguistic cues. A principle bottleneck to further improvement is the shortage of training data (ca. 18k instances in the Penn Discourse Treebank (PDTB)). Shi et al. (2017) proposed to acquire additional data by exploiting connectives in translation: human translators mark discourse relations which are implicit in the source language explicitly in the translation. Using back-translations of such explicitated connectives improves discourse relation parsing performance. This paper addresses the open question of whether the choice of the translation language matters, and whether multiple translations into different languages can be effectively used to improve the quality of the additional data.

1 Introduction

Discourse relations connect two sentences/clauses to each other. The identification of discourse relations is an important step in natural language understanding and is beneficial to various downstream NLP applications such as text summarization (Yoshida et al., 2014; Gerani et al., 2014), question answering (Verberne et al., 2007; Jansen et al., 2014), machine translation (Guzmán et al., 2014; Meyer et al., 2015), and so on.

Discourse relations can be marked explicitly using a discourse connective or discourse adverbial such as "because"", "but", "however", see example 1. Explicitly marked relations are relatively easy to classify automatically (Pitler et al., 2008). In example 2, the causal relation is not marked explicitly, and can only be inferred from the texts. This second type of case is empirically even more common than explicitly marked relations (Prasad et al., 2008), but is much harder to classify automatically.

1. *[No one has worked out the players' average age.]*$_{Arg1}$ **But** *[most appear to be in their late 30s.]*$_{Arg2}$
   — Explicit, Comparison.Contrast

2. *[I want to add one more truck.]*$_{Arg1}$ (**Implicit = Because**) *[I sense that the business will continue grow.]*$_{Arg2}$
   — Implicit, Contingency.Cause

The difficulty in classifying implicit discourse relations stems from the lack of strong indicative cues. Early work has already shown that implicit relations cannot be learned from explicit ones by just removing the discourse markers, which may lead to a meaning shift in the examples (Sporleder and Lascarides, 2008), making human-annotated relations currently the only reliable source for training implicit discourse relation classification.

Due to the limited size of available training data, several approaches have been proposed for acquiring additional training data using automatic methods (Wang et al., 2012; Rutherford and Xue, 2015). The most promising approach so far, Shi et al. (2017), exploits the fact that human translators sometimes insert a connective in their translation even when a relation was implicit in the original text. Using a back-translation method, Shi et al. showed that such instances can be used for acquiring additional labeled text.

Shi et al. (2017) however only used a single target langauge (French), and had no control over the quality of the labels extracted from back-translated connectives. In this paper, we therefore systematically compare the contribution of three target translation languages from different language families: French (a Romance language), German (from the Germanic language family) and Czech (a Slavic language). As all three of these languages are part of the EuroParl corpus, this also allows us to directly test whether higher quality

can be achieved by using those instances that were consistently explicitated in several languages. We use cross-lingual explicitation to acquire more reliable implicit discourse relation instances with separate arguments that are from adjacent sentences in a document, and conducted experiments on PDTB benchmark with multiple conventional settings including cross validation. The experimental results show that the performance has been improved significantly with the additional training data, compared with the baseline systems.

2 Related Work

Recognizing implicit discourse relation, as one of the most important and challenging part of discourse parser system, has drawn a lot of attention in recent years after the release of PDTB (Prasad et al., 2008), the largest available corpus with annotated implicit examples, including two shared task in CoNLL-2015 and CoNLL-2016 (Xue et al., 2015, 2016).

Early attempts focused on statistical machine learning solutions with sparse linguistic features and linear models. They used several linguistically informed features like polarity tags, Levin verb classes and brown cluster etc. (Pitler et al., 2009; Park and Cardie, 2012; Rutherford and Xue, 2014).

Recent methods for discourse relation classification have increasingly relied on neural network architectures (Ji et al., 2016; Qin et al., 2016, 2017; Shi and Demberg, 2018). However, with the high number of parameters to be trained in more and more complicated deep neural network architectures, the demand for more reliable annotated data has become even more urgent. Data extension has been a longstanding goal in implicit discourse relation classification. Wang et al. (2012) proposed to differentiate typical and atypical examples for each relation and augment training data for implicit only by typical explicits. Rutherford and Xue (2015) designed criteria for selecting explicit samples in which connectives can be omitted without changing the interpretation of the discourse. More recently, Shi et al. (2017) proposed a pipeline to automatically label English implicit discourse samples based on explicitation of discourse connectives during human translating in parallel corpora, and achieve substantial improvements in classification. Our work here directly extends theirs by employing document-

aligned cross-lingual parallel corpora and majority votes to get more reliable and in-topic annotated implicit discourse relation instances.

3 Methodology

Our goal here aims at sentence pairs in cross-lingual corpora where connectives have been inserted by human translators during translating from English to several other languages. After back-translating from other languages to English, explicit relations can be easily identified by discourse parser and then original English sentences would be labeled accordingly.

We follow the pipeline proposed in Shi et al. (2017), as illustrated in Figure 1, with the following differences:

- Shi et al. (2017) suffered from the fact that typical sentence-aligned corpora may have some sentences removed and make the sentences no longer coherent to get inter-sentential discourse relation instances. Here we filter and re-paragraph the line-aligned corpus to parallel document-aligned files, which makes it possible to obtain in-topic inter-sentential instances. After preprocessing, we got 532,542 parallel sentence pairs in 6,105 documents.

- Shi et al. (2017) pointed out that having correct translation of explicit discourse connective is more important than having the correct translation of the whole sentence. In this paper we use a statistical machine translation system instead of a neural one for more stable translations of discourse connectives.

- Instead of a single language pair, we use three language pairs and majority votes between them to get annotated implicit discourse relation instances with high confidence.

Figure 1 illustrates the pipeline of our approach. It consists of a few steps including preprocessing, back-translating, discourse parsing and majority voting. For each document, we back-translate its German, French and Czech translation back to English with the MT system and parse them with discourse parser. In this way, we can easily identify those instances that are originally implicit but explicit in German, French or Czech. With majority vote by the explicit examples in those three languages, the original English instance could be labeled with different confidences.
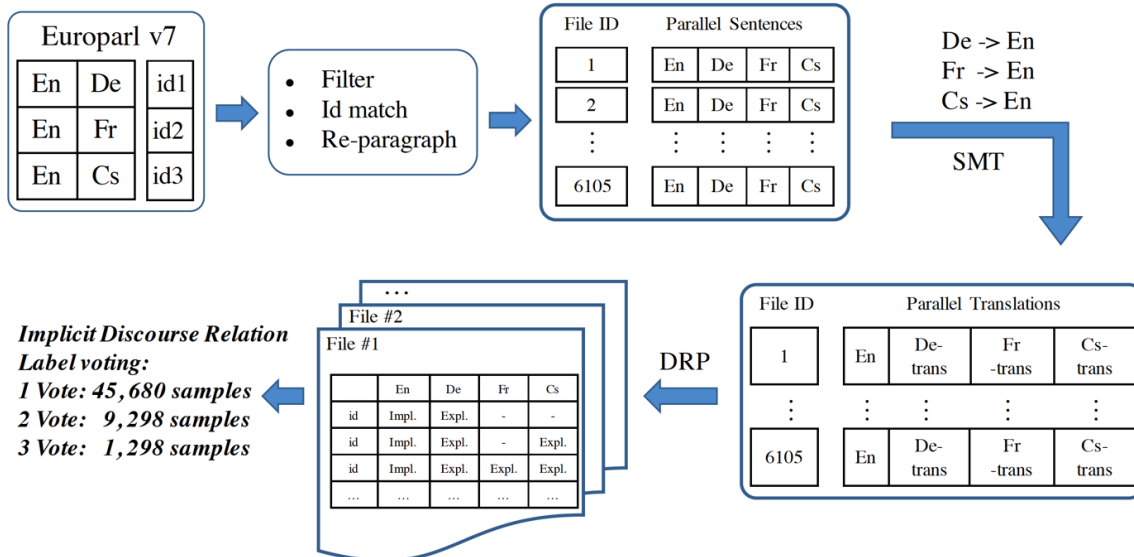
Figure 1: The pipeline of proposed method. "SMT" and "DRP" denote statistical machine translation and discourse relation parser respectively.

3.1 Preprocessing

We use European Parliament Proceedings Parallel Corpus (Europarl[1]) (Koehn, 2005) and choose English-French, German and Czech pairs as our parallel corpora. Each source-target pair consists of source and target sentences along with a sentence ID with which we could easily identify the location of the sentence in certain paragraphs. In order to get document-aligned parallel sentences among all these four languages, we do preprocessing steps as follows:

- Filtering: remove those sentences that don't have all the three translations in French, German or Czech.

- ID matching: re-group each sentence into different documents by the sentence IDs.

- Re-paragraph: rank the sentences in each documents by the ID and re-paragraph them.

3.2 Machine Translation

We train three MT systems to back-translate French, German and Czech to English. To have word alignments, better and stable back-translations, we employ a statistical machine translation system MOSES (Koehn et al., 2007), trained on the same parallel corpora. Source and target sentences are first tokenized, true-cased and then fed into the system for training. In our case,

the translation target texts are identical with the training set of the translation systems; this would not be a problem because our only objective in the translation is to back-translate connectives in the translation into English. On the training set, the translation system achieves BLEU scores of 66.20 (French), 65.30 (German) and 69.05 (Czech).

3.3 Discourse Parser

We employ the PDTB-style parser proposed in (Lin et al., 2014), which achieved about 96% accuracy on explicit connective identification, to pick up those explicit examples in back-translations in each document. Following the definitions of discourse relations in the PDTB that the arguments of the implicit discourse relations should be adjacent sentences but not for the explicit relations, we screen out all those explicit samples from the outputs of the parser that don't have consecutive arguments.

3.4 Majority Vote

After parsing the back-translations of French, German and Czech, we can compare whether they contain explicit relations which connect the same relational arguments. The analysis of this subset then allows us to identify those instances that could be labeled with high confidence, i.e. where back-translations from all three languages allow us to infer the same coherence label. Note that it is not necessarily the case that all back-translations contain an explicitation for the same instance (for

instance, the French translator may have explicitated a relation, while the German and the Czech translators didn't do so), or that they propose *the same* coherence label: the human translation can introduce "noise" in the sense of the human translators inferring different coherence relations, the machine translation model can introduce errors in back-translation, and the discourse parser can mislabel ambiguous explicit connectives. When we use back-translations of several languages, the idea is that we can eliminate much of this noise by selecting only those instances where all back-translations agree with one another, or the ones where at least two back-translations allow us to infer identical labels.

Figure 2 illustrates the number of automatically labeled implicit discourse relation examples together with the information of how many of the instances that just one, two or all three back-translations provided the same labels.

In the One Vote agreement, every explicit relation has been accepted and the original implicit English sentences have been annotated correspondingly. Likewise, Two Votes agreement needs at least two out of three languages to have the same explicit relation label after back-translation; agreement between all three back-translations is denoted as Three Votes.

4 Experiments and Results

4.1 Data

Europarl Corpora: The parallel corpora used here are from Europarl (Koehn, 2005), it contains about 2.05M English-French, 1.96M English-German and 0.65M English-Czech pairs. After preprocessing, we got about 0.53M parallel sentence pairs in all these four languages.

The Penn Discourse Treebank (PDTB): PDTB (Prasad et al., 2008) is the largest available manually annotated corpus of discourse relations from Wall Street Journal. Each discourse relation has been annotated in three hierarchy levels. In this paper, we follow the previous conventional settings and focus on the second-level 11-ways classification (Lin et al., 2009; Ji and Eisenstein, 2015; Rutherford et al., 2017; Shi et al., 2017), after removing the relations with few instances.

4.2 Implicit discourse relation classification

To evaluate whether the extracted data is helpful to this task, we use a simple and effective
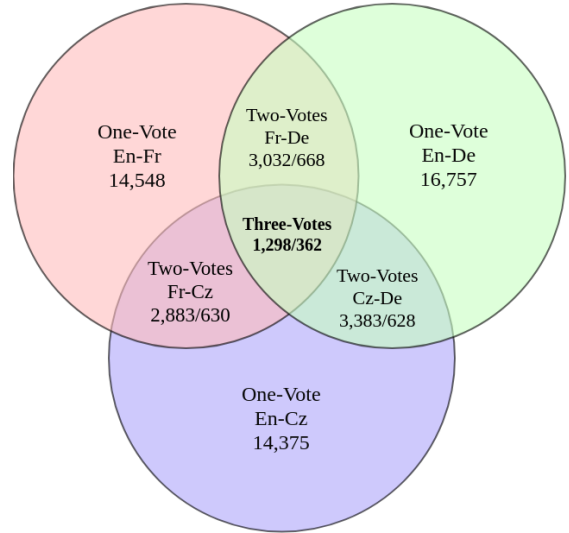

Figure 2: Numbers of implicit discourse relation instances from different agreements of explicit instances in three back-translations. En-Fr denotes instances that are implicit in English but explicit in back-translation of French, same for En-De and En-Cz. The overlap means they share the same relational arguments. The numbers under "Two-Votes" and "Three-Votes" are the numbers of discourse relation agreement / disagreement between explicits in back-translations of two or three languages.
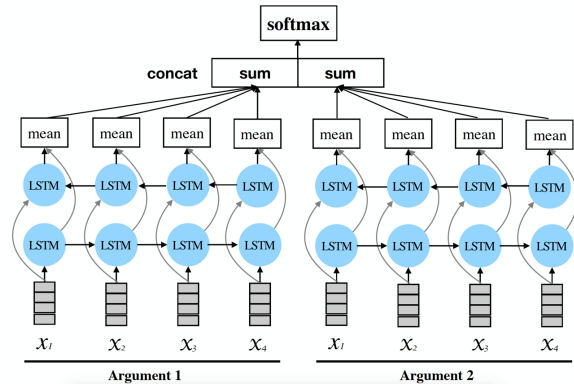

Figure 3: Bi-LSTM network for implicit discourse relation classification.

bidirectional Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network.

A LSTM recurrent neural network processes a variable-length sequence $x = (x_1, x_2, ..., x_n)$. At time step t, the state of memory cell c_t and hidden h_t are calculated with the Equations 1:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ \hat{c}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} W \cdot [h_{t-1}, x_t] \quad (1)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t$$
$$h_t = o_t \odot \tanh(c_t)$$

After being mapped to vectors, words are fed into the network sequentially. Hidden states of LSTM cell from different directions are averaged. The representations of two arguments from two separate bi-LSTMs are concatenated before being fed into a softmax layer for prediction. The architecture is illustrated in Figure 3.

Implementation: The model is implemented in Pytorch[2]. All the parameters are initialized uniformly at random. We employ cross-entropy as our cost function, Adagrad with learning rate of 0.01 as the optimization algorithm and set the dropout layers after embedding and output layer with drop rates of 0.5 and 0.2 respectively. The word vectors are pre-trained word embeddings from Word2Vec[3].

Settings: We follow the previous works and evaluate our data on second-level 11-ways classification on PDTB with 3 settings: Lin et al. (2009) (denotes as PDTB-Lin) uses sections 2-21, 22 and 23 as train, dev and test set; Ji and Eisenstein (2015) uses sections 2-20, 0-1 and 21-22 as train, dev and test set; Moreover, we also use 10-folds cross validation among sections 0-23 (Shi and Demberg, 2017). For each experiment, the additional data is only added into the training set.
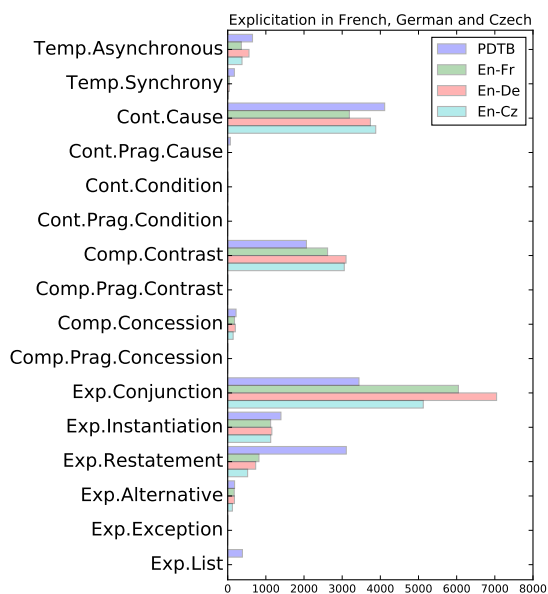

Figure 4: Distributions of PDTB and the extracted data among each discourse relation.

4.3 Results

4.3.1 Distribution of new instances

Figure 4 shows the distributions of expert-annotated PDTB implicit relations and the implicit discourse examples extracted from the French, German and Czech back-translations. Overall, there is no strong bias – all relations seem to be represented similarly well, in line with their general frequency of occurrence. One interesting exception is the higher number of *Expansion.Conjunction* relation from the German translations. The over-representation of *Expansion.Conjunction* relation in German indicates that German translators tend to use more explicit cues to mark these relations. This is an independently discovered well-known finding from the literature (Kunz and Lapshinova-Koltunski, 2015), which observed that German tends to mark conjunction relations with discourse cues, while English tends to use coreference instead. We also find that *Expansion.Restatement* relations are under-represented in our back-translation method, indicating that these relations are explicitated particularly rarely in translation. We also find that we can identify more *Contingency.Cause* and *Comparison.Contrast* relations from the German and Czech back-translations compared to the French ones. This provides us with an interesting lead for future work, to investigate whether French tends to explicitate these relations less, expressing them implicitly like in the English original, or whether French connectives for causal and contrastive relations are more ambiguous, causing problems in the back-translations.

Figure 5 shows that the filtering by majority votes (including only two cases where at least two back-translations agree with one another vs. where all three agree) does again not change the distribution of extracted relations.

In summary, we can conclude that the choice of translation language *can* matter: depending on what types of relations are most important to acquire more data for the target task at hand, a language that tends to explicitate that relation frequently can be particularly suitable. On the other hand, if no strong such preferences on labelling specific relations exist, we can see that the choice of translation language only has a minor effect on the overall distribution of additional implicit discourse relation labels.
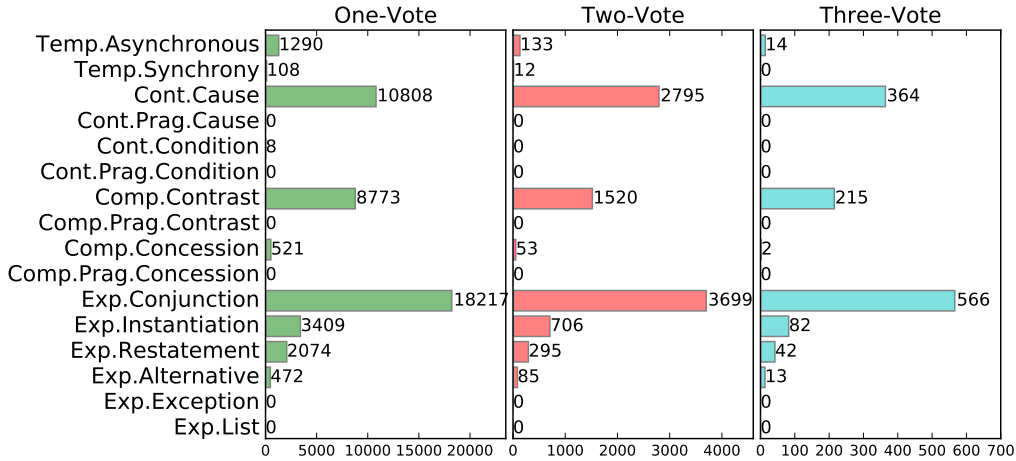
Figure 5: Distributions of discourse relations with different agreements.

	PDTB-Lin	PDTB-Ji	Cross Validation	size of extra data
Majority Class	26.11	26.18	25.59	-
Rutherford et al. (2017)	38.38	-	-	-
Shi et al. (2017)	**45.50**	-	37.84	102,314
PDTB only	37.95(0.59)	40.57(0.67)	37.82(0.14)	-
PDTB + En-Fr	38.96(0.69)	40.14(0.78)	38.32(0.62)	14,548
En-De	39.65(0.95)	39.96(0.44)	37.97(0.46)	16,757
En-Cz	37.90(1.27)	40.59(0.51)	37.42(0.50)	14,375
All	37.73(0.74)	40.41(0.65)	37.16(0.64)	45,680
PDTB + 2-votes	**40.34**(0.75)	**41.95**(0.97)	**38.98**(0.14)	9,298
PDTB + 3-votes	39.88(0.79)	41.19(0.63)	38.33(0.50)	1,298

Table 1: Performances with different sets of additional data. Average accuracy of 10 runs (5 for cross validations) are shown here with standard deviation in the brackets. Numbers in bold are significantly ($p<0.05$) better than the *PDTB only* baseline with unpaired t-test.

4.3.2 Quantitative Results

Table 1 shows that best results are achieved by adding only those samples for which two back-translations agree with one another. This may represent the best trade-off between reliability of the label and the amount of additional data. The setting where the data from all languages is added performs badly despite the large number of samples, because this method contains different labels for the same argument pairs, for all those instances where the back-translations don't yield the same label, thus introducing noise into the system. The size of the extra data used in Shi et al. (2017) is about 10 times larger than our 2-votes data. The selection of instances differs in their paper from ours, in that they only use French, and in that they, unlike this paper, focus on intra-sentential samples. The model using the few reliable samples extracted from the back-translations of the three languages here significantly outperforms the results found in Shi et al. (2017) in the cross-validation setting. On the PDTB-Lin test set, we don't match performance, but note that this test set is based only on 800 instances, as opposed to the 16k instances in the cross-validation evaluation.

4.3.3 Qualitative analysis

Finally, we want to provide insight into what kind of instances the system extracts, and why back-translation labels sometimes disagree. We have identified four major cases based on a manual analysis of 100 randomly sampled instances.

Case 1: Sometimes, back-translations from several languages may yield the same connective because the original English sentence actually was not really unmarked, but rather contained an expression which could not be automatically recog-

17

nized as a discourse relation marker by the automatic discourse parser[4]. This can actually help us to identify new alternative lexicalisation for discourse relations, and thus represents a promising technique for improving discourse relation classification also on texts for which no translations are available.

Original English: I presided over a region crossed by heavy traffic from all over Europe, with significant accidents which gave rise to legal actions. *What is more,* In 2002, two Member States of the European Union appealed to the European Court of Justice to repeal Directive 2002/15/EC because it included self-employed drivers ; the Court rejected their appeal on the grounds of road safety.

French back-translation: I presided over a region crossed by heavy traffic from the whole of Europe, with significant accidents which gave rise to legal actions, **moreover**, (Expansion.Conjunction) in 2002 , two Member States have appeal on the European Court of Justice, which has condemned the rejection of the grounds of road safety.

German back-translation: I presided over a region crossed by heavy traffic from across Europe, with significant accidents which, **moreover** (Expansion.Conjunction) in 2002, two Member States of the European Union appealed to the European Court of Justice to repeal Directive 2002/15/EC , because it included self-employed drivers ; the Court quashed for reasons of road safety.

Czech back-translation: I was in the region with very heavy traffic from all over Europe, with significant accidents which gave rise to legal actions **therefore** (Contingency.Cause) after all, in 2002, two Member States of the European Union appealed to the European Court of Justice to repeal Directive 2002/15/EC that also applies to self-employed drivers; the Court rejected their appeal on the grounds of road safety.

The expression *what is more* is not part of the set of connectives labeled in PDTB and hence was not identified by the discourse parser. Our method is successful because such cues can be automatically identified from the consistent back-translations into two languages. (The case in Czech is more complex because the back-translation contains two signals, *therefore* and *after all*, see case 4.)

We also found some similar expressions in this case like:

"in reality" ("implicit", original English) → "in fact" (explicit, back-translation);

"for that reason" → "therefore";

"this is why" → "therefore";

"be that as it may" → "however / nevertheless";

"for another" → "furthermore / on the other hand";

"in spite of that" → "however / nevertheless" and so on.

Case 2: Majority votes help to reduce noise related to errors introduced by the automatic pipeline, such as argument or connective misidentification: in the below example, *also* in the French translation is actually the translation of *along with*.

Original English: on behalf of the PPE-DE Group. (DE) Madam President, Commissioner, ladies and gentlemen, the public should be able to benefit in two ways from the potential for greater road safety. *For this reason*, along with the report we are discussing today, I call for more research into the safety benefits of driver-assistance systems.

French back-translation: (DE) Madam President, Commissioner, ladies and gentlemen, citizens should be able to benefit in two ways of the possibility of improving road safety. **also** (Expansion.Conjunction) when we are discussing this report today, I appeal to the intensification of research at the level of the benefits of driver-assistance systems in terms of security, as well as the transmission of information about them.

German back-translation: (DE) Madam President, Commissioner, ladies and gentlemen, road safety potentials should citizens in the dual sense **therefore** (Contingency.Cause) I urge, together with the report under discussion today, the prevention and education about the safety benefits of driver-assistance systems.

Czech back-translation: (DE) Madam President, Commissioner, ladies and gentlemen, the public would be the potential for greater road safety should have a two-fold benefit, **therefore** (Contingency.Cause) I call, in addition to the report, which we are debating today , for more research and education in the safety benefits of driver-assistance systems.

Case 3: Discrepancies between connectives in back-translations can also be due to differences in how translators interpreted the original text. Here are cases of genuine ambiguities in the implicit discourse relation.

Original English: with regard, once again, to European Union law, we are dealing in this case with the domestic legal system of the Member States. *That being said*, I cannot answer for the Council of Europe or for the European Court of Human Rights, which have issued a decision that I understand may raise some issues for Parliament.

---

[4]In the following examples, the original English sentence is shown is followed by the back-translations from French, German and Czech along with the connectives and senses.

French back-translation: with regard, once again, the right of the European Union, we are here in the domestic legal system of the Member States. **however,** (Comparison.Contrast) I cannot respond to the place of the Council of Europe or for the European Court of Human Rights, which have issued a decision that I understand may raise questions in this House.

German back-translation: once again on the right of the European Union, we have it in this case with the national legal systems of the Member States. **therefore,** (Contingency.Cause) I cannot, for the Council of Europe and the European Court of Human Rights, which have issued a decision, which I can understand, in Parliament raises some issues.

Czech back-translation: I repeat that, when it comes to the European Union, in this case we are dealing with the domestic legal system of the Member States. **in addition,** (Expansion.Conjunction) I cannot answer for the Council of Europe or for the European Court of Human Rights , which has issued a decision that I understand may cause in Parliament some doubts.

Case 4: Implicit relations can co-occur with marked discourse relations (Rohde et al., 2015), and multiple translations help discover these instances, for example:

Original English: We all understand that nobody can return Russia to the path of freedom and democracy, *(implicit: but) what is more*, the situation in our country is not as straightforward as it might appear to the superficial observer.

French back-translation: we all understand that nobody can return Russia on the path of freedom and democracy but Russia itself, its citizens and its civil society **but** (Comparison.Contrast) there is more, the situation in our country is not as simple as it might appear to be a superficial observer.

German back-translation: we are all aware that nobody Russia back on the path of freedom and democracy, as the country itself, its people and its civil society **but** (Comparison.Contrast) the situation in our country is not as straightforward as it might appear to the superficial observer.

Czech back-translation: we all know that Russia cannot return to the path of freedom and democracy there, but Russia itself, its people and civil society. **in addition** (Expansion.Conjunction) the situation in our country is not as straightforward as it might appear to the superficial observer.

5   Conclusion

We compare the explicitations obtained from translations into three different languages, and find that instances where at least two back-translations agree yield the best quality, significantly outperforming a version of the model that does not use additional data, or uses data from just one language.

We also found that specific properties of the translation language affect the distribution of the additionally acquired data across coherence relations: German, for instance, is known to mark conjunction relations using discourse cues more frequently, while English and other languages tend to express these relations rather through lexical cohesion or pronouns. This was reflected in our experiments: we found a larger proportion of explicitations for conjunction relations in German than the other translation languages.

Finally, our qualitative analysis shows that the strength of the method partially stems from being able to learn additional discourse relation signals because these are typically translated consistently. The method thus shows promise for the identification of discourse markers and alternative lexicalisations, which can subsequently be exploited also for discourse relation classification in the absence of translation data. Our analysis also shows that our method is useful for identifying cases where multiple relations holding between two arguments.

6   Acknowledgments

References

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 687–698.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 977–986.

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.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 332–342. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Kerstin Kunz and Ekaterina Lapshinova-Koltunski. 2015. Cross-linguistic analysis of discourse variation across registers. *Nordic Journal of English Studies*, 14(1):258–288.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore. Association for Computational Linguistics.

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

Thomas Meyer, Najeh Hajlaoui, and Andrei Popescu-Belis. 2015. Disambiguating discourse connectives for statistical machine translation. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(7):1184–1197.

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 47th Annual Meeting of the Association for Computational Linguistics*, pages 683–691, Suntec, Singapore. Association for Computational Linguistics.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. *Technical Reports (CIS)*, page 884.

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*, Marrakech, Morocco. European Language Resources Association.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Implicit discourse relation recognition with context-aware character-enhanced embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P. Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.

Hannah Rohde, Anna Dickinson, Chris Clark, Annie Louis, and Bonnie Webber. 2015. Recovering discourse relations: Varying influence of discourse adverbials. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 22–31.

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*, volume 1, 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. 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 on Human Language Technology*, pages 799–808. Association for Computational Linguistics.

20

Wei Shi and Vera Demberg. 2017. On the need of cross validation for discourse relation classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 150–156. Association for Computational Linguistics.

Wei Shi and Vera Demberg. 2018. Learning to explicitate connectives with seq2seq network for implicit discourse relation classification. *arXiv preprint arXiv:1811.01697*.

Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. 2017. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 484–495.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–736. ACM.

Xun Wang, Sujian Li, Jiwei Li, and Wenjie Li. 2012. Implicit discourse relation recognition by selecting typical training examples. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING-2012)*, pages 2757–2772.

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 CoNLL-15 Shared Task*, pages 1–16. Association for Computational Linguistics.

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

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1839.