Using Explicit Discourse Connectives in Translation for Implicit Discourse Relation Classification

Wei Shi[1], **Frances Yung**[1], **Raphael Rubino**[1,3] and **Vera Demberg**[1,2]
[1]Dept. of Language Science and Technology
[2]Dept. of Mathematics and Computer Science, Saarland University
[3]German Research Center for Artificial Intelligence (DFKI)
Saarland Informatic Campus, 66123 Saarbrücken, Germany
{w.shi, frances, vera}@coli.uni-saarland.de
raphael.rubino@dfki.de

Abstract

Implicit discourse relation recognition is an extremely challenging task due to the lack of indicative connectives. Various neural network architectures have been proposed for this task recently, but most of them suffer from the shortage of labeled data. In this paper, we address this problem by procuring additional training data from parallel corpora: When humans translate a text, they sometimes add connectives (a process known as *explicitation*). We automatically back-translate it into an English connective, and use it to infer a label with high confidence. We show that a training set several times larger than the original training set can be generated this way. With the extra labeled instances, we show that even a simple bidirectional Long Short-Term Memory Network can outperform the current state-of-the-art.

1 Introduction

When humans comprehend language, their interpretation consists of more than just the sum of the content of the sentences. Additional semantic relations (known as coherence relations or discourse relations) are inferred between sentences in the text. Identification of discourse relations is useful for various NLP applications such as question answering (Jansen et al., 2014; Liakata et al., 2013), summarization (Maskey and Hirschberg, 2005; Yoshida et al., 2014; Gerani et al., 2014), machine translation (Guzmán et al., 2014; Meyer et al., 2015) and information extraction (Cimiano et al., 2005). Recently, the task has drawn increasing attention, including two CoNLL shared tasks (Xue et al., 2015, 2016).

Discourse relations are sometimes expressed with an *explicit* discourse connective (DC), such as "because", "but", "if". Example 1 shows an explicit discourse relation marked by "because"; the text spans between which the relation holds are marked as *Arg1* and *Arg2*. DCs serve as strong cues and allow us to classify discourse relations with high accuracy (Pitler et al., 2008, 2009; Lin et al., 2014).

However, more than half of the discourse relations in a text are not signalled by a connective. See for example 2: a contrastive relation can be inferred between the text spans marked as *Arg1* and *Arg2*. Implicit relation classification is very challenging and represents a bottleneck of the entire discourse parsing system.

1. *[The city's Campaign Finance Board has refused to pay Mr Dinkins $95,142 in matching funds]*$_{Arg1}$ ***because*** *[his campaign records are incomplete.]*$_{Arg2}$
   — *Explicit, Contingency.Cause*

2. *[They desperately needed somebody who showed they cared for them, who loved them.]*$_{Arg1}$ *[The last thing they needed was another drag-down blow.]*$_{Arg2}$
   —*Implicit, Comparison.Contrast*

In order to classify an implicit discourse relation, it is necessary to represent the semantic content of the relational arguments, which may give a cue to the coherence relation, e.g. "care" – "drag-down blow" in 2. Early methods have focused on designing various features to overcome data sparsity and more effectively identify relevant concepts in the two discourse relational arguments. (Lin et al., 2009; Zhou et al., 2010; Biran and McKeown, 2013; Park and Cardie, 2012; Rutherford and Xue, 2014), while recent efforts use distributed representations with neural network architectures (Zhang et al., 2015; Ji and Eisenstein,

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 484–495,
Taipei, Taiwan, November 27 – December 1, 2017

2015; Ji et al., 2016; Chen et al., 2016; Qin et al., 2016, 2017). Both streams of methods suffer from insufficient annotated data (Wang et al., 2015), since the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), which is the discourse annotated resource mostly used by the community, consists of just 12763 implicit instances in the usual training set and 761 relations in the test set. Some second-level relations only have about a dozen instances. It is therefore crucial to obtain extra data for machine learning.

In this paper, we propose a simple approach to automatically extract samples of implicit discourse relations from parallel corpus via back-translation: Our approach is motivated by the fact that humans sometimes omit connectives during translation (*implicitation*), or insert connectives not originally present in the source text (*explicitation*) (Laali and Kosseim, 2014; Koppel and Ordan, 2011; Cartoni et al., 2011; Hoek and Zufferey, 2015; Zufferey, 2016). When explicitating an implicit relation, the human translator is, in other words, disambiguating the source implicit relation with an explicit DC in the target language.

Our contribution is twofold: Firstly, we propose a pipeline to automatically label English implicit discourse relation samples based on explicitation of DCs in human translation, which is the target side of a parallel corpus. Secondly, we show that the extra instances mined by the proposed method improve the performance of a standard neural classifier by a large margin, when evaluated on the PDTB 2.0 benchmark test set as well as by cross-validation (Shi and Demberg, 2017).

2 Related Work

Early works addressing discourse relation parsing were trying to classify unmarked discourse relations by training on explicit discourse relations with the marker been removed (Marcu and Echihabi, 2002). While this method promised to provide almost unlimited training data, it was shown that explicit relations differ in systematic ways from implicit relations (Asr and Demberg, 2012), so that performance on implicits is very poor when learning on explicits only (Sporleder and Lascarides, 2008).

The release of PDTB (Prasad et al., 2008), the largest available corpus which annotates implicit examples, lead to substantial improvements in classification of implicit relations, and spurred a variety of approaches to the task, including feature-based methods (Pitler et al., 2009; Lin et al., 2009; Park and Cardie, 2012; Biran and McKeown, 2013; Rutherford and Xue, 2014) and neural network models (Zhang et al., 2015; Ji and Eisenstein, 2015; Ji et al., 2016; Chen et al., 2016; Qin et al., 2016, 2017). However, the limited size of the annotated corpus, in combination with the difficulty of the task of inferring the type of relation between given text spans, presents a problem both in training (Rutherford et al. (2017) find that a simple feed-forward architecture can outperform more complex architectures, and argues that the larger number of parameters can not be estimated adequately on the small amount of training data) and testing (Shi and Demberg (2017) report experiments showing that results on the standard test set are not reliable due to the small set of just 761 relations).

Data extension has therefore been a longstanding goal in discourse relation classification. The main idea has been to select explicit discourse instances that are similar to implicit ones to add to the training set. Wang et al. (2012) proposed to differentiate typical and atypical examples for each discourse relation, and augment training data for implicits only by typical explicits. In a similar vein, Rutherford and Xue (2015) proposed criteria for selecting among explicitly marked relations ones that contain discourse connectives which can be omitted without changing the interpretation of the discourse. These relations are then added to the implicit instances in training.

On the other hand, Lan et al. (2013) presented multi-task learning based systems, which in addition to the main implicit relation classification task, contain the task of predicting previously removed connectives for explicit relations, and profit from shared representations between the tasks. Similarly, Hernault et al. (2010) observes features that occur in both implicit and explicit discourse relations, and exploit such feature co-occurrence to extend the features for classifying implicits using explicitly marked relations. Mihăilă and Ananiadou (2014) and Hidey and McKeown (2016) proposed semi-supervised learning and self-learning methods to improve recognition of patterns that typically signal causal discourse relations.

The approach proposed here differs from previous approaches, because we extend our train-

ing data only by originally implicit relations, and obtain the label through the disambiguation that sometimes happens in human translation.

Parallel corpora have been exploited as a resource of discourse relation data in previous work but have mostly been used with goals different from ours: Cartoni et al. (2013) and Meyer et al. (2015) use parallel corpora to label and disambiguate discourse connectives in the target language based on explicitly marked English relations, in order to help machine translation. A second application has been to project discourse annotation from English onto other languages through parallel corpora, in order to construct discourse annotated resources for the target language (Versley, 2010; Zhou et al., 2012; Laali and Kosseim, 2014).

The approach that is in spirit most similar to ours is by Wu et al. (2016), who extracted bilingual-constrained synthetic implicit data from a sentence-aligned English-Chinese corpus and got improvements by incorporating these data via a multi-task neural network on the 4-way classification.

3 Method

Our proposed method aims at sentence pairs in the parallel corpora where an *implicit* discourse relations on the source English side has been translated by human translators into an explicitly marked relation on the target side. The inserted connective hence disambiguates the originally implicit relation, and the discourse relation can be classified with confidence (under the assumption that the same discourse relation holds in the original source text).

The pipeline of our approach is detailed in below steps.

1. The target side of a sentence-aligned parallel corpus, with English as the source text, is back-translated to English using a pre-trained machine translation system.
2. An end-to-end discourse relation parser for English is run on both the source side and the back-translated target side. The parser will output a list of explicit and implicit relations, including the relation sense and argument spans of each relation.
3. Implicit-to-explicit discourse relation alignments are identified according to the output of the end-to-end parser. Implicit relations in the PDTB are only ever annotated between consecutive sentences. Therefore, we specifically extract pairs of consecutive sentences on the source English side:
   - that are identified as the *Arg1* and *Arg2* of an *implicit* discourse relation[1];
   - whose corresponding back-translated target sentences are identified as the *Arg1* and *Arg2* of an *explicit* relation;
   - that are not part of the *Arg1* or *Arg2* of any other discourse relations[2].
4. Label the source English implicit relation with the relation class of the explicit relation in back-translated target text. The two consecutive sentences are marked as *Arg1* and *Arg2* respectively.

Figure 1 illustrates the pipeline of our approach, which takes an English-to-French parallel corpus as input and outputs a list of implicit discourse relations, each containing two arguments from the source English text and a relation class according to the back-translated French DC.

We then compare the performance of a neural implicit discourse relation classifier trained with the annotated implicit relation samples in PDTB alone and also with the extra training samples mined from the parallel corpus. The classifier performance is evaluated on the standard PDTB implicit relation test set and by cross-validation.

3.1 Advantages of using back-translation

In the proposed method, we disambiguate implicit relations according to the explicitated translation. Instead of directly classifying the explicit relation in the target language, we back-translate the target text to the source language by machine translation (MT) because:

- Discourse parsers on low-resource languages do not perform well, or are even not available.
- Different languages have different sets of discourse relation classes defined. By the means of back-translation, we can use an English discourse parser on the target text, and thus

[1] Relations signaled by *Alternative Lexicalization* are counted as implicit relations and extracted as samples. However, *NoRel* and *EntRel* are excluded.

[2] This restriction avoids mis-alignment of relations between source and target texts.

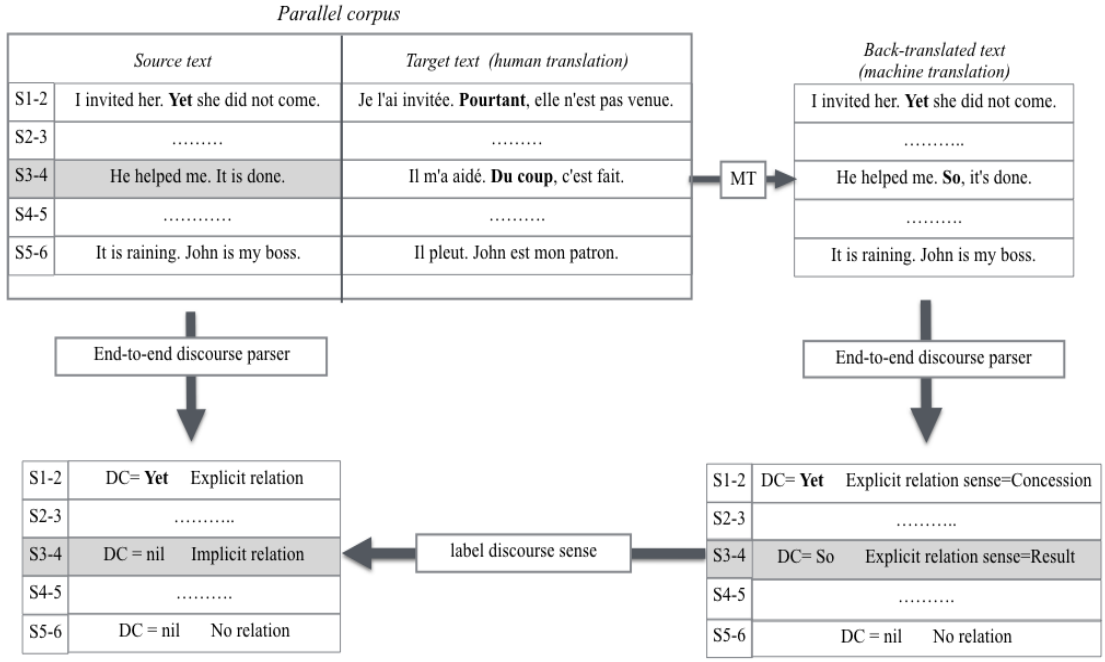

Figure 1: Pipeline showing how an implicit discourse relation sample, sentence pair 3-4, is extracted and labeled using a parallel corpus.

label the implicit relations with the same set of relation labels defined for English.

- The quality of the MT system has limited impact on our approach. Since the DC tokens are powerful features to disambiguate an explicit relation, limited contextual features are required. We just need correct translation of the explicit DC tokens, irrespective of word order and the rest of the translation.

3.2 Inter-sentential and intra-sentential relations

Only inter-sentential implicit relations are annotated in the PDTB, due to time and resource constraints (Prasad et al., 2008). However, this does not mean that implicit relations only hold between consecutive sentences.

We decided to extract intra-sentential relation samples from the parallel corpus based on two motivations: Firstly, we hypothesize that intra-sentential implicit relations share similar features as inter-sentential ones. Including both types may hence increase dataset size. In fact, we will see in the experiment results that intra-sentential training samples largely improve classification of implicit relations, even though the test data from PDTB contains inter-sentential samples only. An analysis on what we learn from the intra-sentential samples is presented in Section 6.1.

Secondly, intra-sentential relations can potentially be identified with higher reliability: Parallel corpora are typically sentence-aligned. This makes it a lot easier to extract sentences that are detected by the end-to-end discourse relation parser as explicit in the (back-)translation target side but not on the original source side, without needing to worry about whether any sentences in the dataset were removed or the order changed during preprocessing (which would be detrimental for detecting intra-sentential relations).

3.3 Argument spans

It is possible but not entirely trivial to determine the argument spans of the discourse relations labeled with the back-translation method. In this paper, we chose a neural network model that concatenates the *Arg1* and *Arg2* representations (see Section 4.4), so that determining exact text spans of Arg1 and Arg2 was not necessary. We are not the first one to do like this, in the work by Rönnqvist et al. (2017), they modeled the *Arg1-Arg2* pairs as a joint sequence and did not compute intermediate representations of arguments separately, to make it more generally flexible in modeling discourse units and easily extend to additional contexts.

4 Experiment

4.1 Data

Parallel Corpora The corpora used for the extraction of implicit discourse relation samples are publicly available bilingual English-French parallel datasets compiled by Rabinovich et al. (2015).[3] They consist of European parliamentary proceedings, literary works and the Hansard corpus – genres that are different from the PDTB, because we want to expand the diversity of discourse relation samples available in the PDTB. These corpora contain a total of $\sim$ 1.9M sentence pairs with an average of 22.7 words per English sentence. Each corpus contains an originally written part in English (used as target for the MT system) and its corresponding human translation in French (used as source). We use the same corpora to train the French–English MT system (Section 4.2), to back-translate the French side into English and to extract additional discourse training data.

The Penn Discourse Treebank (PDTB) We use the Penn Discourse Treebank 2.0 (Prasad et al., 2008) for the training and testing of the implicit discourse relation classifier. PDTB is the largest available manually annotated corpus of explicit and implicit discourse relations based on one million word tokens from the Wall Street Journal. Each discourse relation is annotated with at most two senses from a three-level hierarchy of discourse relations. The first level roughly categorizes the relations into four major classes, each of which is further categorized in to more distinct relation types. Conventionally, discourse relation classifiers are either evaluated by the accuracy of the first-level 4-way classification(Pitler et al., 2009; Rutherford and Xue, 2014; Chen et al., 2016), or the second-level 11-way classification (Lin et al., 2009; Ji and Eisenstein, 2015; Qin et al., 2016, 2017).

4.2 Machine Translation System

We train an MT system to back-translate the target side of the parallel corpus to English. To produce the highest-quality back-translation, we use a neural MT system trained on the same parallel corpus. The system is implemented by Open-source Neural Machine Translation (OpenNMT) (Klein et al., 2017). Source words are first mapped to word vectors and then fed into a recurrent neural network. At each target time step, *attention* is applied over the source RNN and combined with the current hidden state to produce a prediction of the next word, and this prediction would be fed back into the target RNN.

We evaluate the MT system on *newstest2014* and *newsdiscusstest2015*, reaching 24.63 and 22.58 BLEU respectively. The French side of the training data back-translated into English is evaluated against the originally written English source, leading to a BLEU score of 34.17.[4] The evaluation of the back-translated corpus indicates that the source text is not exactly reproduced. Critically, we assume that the MT system preserves the explicitness of the target DCs, instead of explicitating or implicitating DCs as in the human translation.

4.3 End-to-end discourse parser

We employ the PDTB-style End-to-End Discourse Parser (Lin et al., 2014) to identify and classify the explicit instances from the back-translated English sentences. It achieved about 87% F1 score for explicit relations on level-2 types, even higher than human agreement of 84%. The accuracy on explicit DC identification is 96%.

On the source side, the end-to-end parser is applied to pick implicit relations from other types of relations, i.e. explicit relations or *no relation*, in order to extract implicit-to-explicit DC translation from the parallel corpus[5]. On the back-translation, the end-to-end parser is applied to identify only explicitly marked discourse relations.

4.4 Implicit relation classification model

We use a Bidirectional Long Short-Term Memory (LSTM) network as the implicit relation classification model to evaluate the samples extracted by the proposed method. This architecture inspects both left and right contextual information and has been proven effective in relation classification (Zhou et al., 2016; Rönnqvist et al., 2017).

The model is illustrated in Figure 2, where each word from the two discourse relational arguments is represented as a vector, which is found through a look-up word embedding. Given the word representations $[\mathbf{w}_1,\mathbf{w}_2,...,\mathbf{w}_n]$ as the input sequence, an

[3] All corpora are available at http://cl.haifa.ac.il/projects/translationese/

[4] Case sensitive BLEU implemented in *mteval-v13a.pl*. Test sets available at http://www.statmt.org/wmt15/translation-task.html

[5] The non-explicit sense classification module of this parser is thus not used in the proposed method.

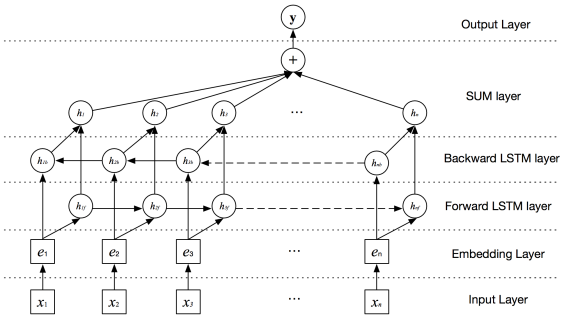

Figure 2: The bidirectional LSTM Network for the task of implicit discourse relation classification.

LSTM computes the state sequence [$\mathbf{h}_1$,$\mathbf{h}_2$,...,$\mathbf{h}_n$] with the following equations:

$$\mathbf{i}_t = \sigma(\mathbf{W}_w^i \mathbf{w}_t + \mathbf{W}_h^i \mathbf{h}_{t-1} + \mathbf{W}_c^i \mathbf{w}_{t-1} + \mathbf{b}_i)$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_w^f \mathbf{w}_t + \mathbf{W}_h^f \mathbf{h}_{t-1} + \mathbf{W}_c^f \mathbf{w}_{t-1} + \mathbf{b}_f)$$
$$\mathbf{g}_t = tanh(\mathbf{W}_w^c \mathbf{w}_t + \mathbf{W}_h^c \mathbf{h}_{t-1} + \mathbf{b}_c)$$
$$\mathbf{c}_t = \mathbf{f}_t \bigodot \mathbf{c}_{t-1} + \mathbf{i}_t \bigodot \mathbf{g}_t$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_w^o \mathbf{w}_t + \mathbf{W}_h^o \mathbf{h}_{t-1} + \mathbf{b}_o)$$
$$\mathbf{h}_t = tanh(\mathbf{c}_t) \bigodot \mathbf{o}_t$$

The forward and backward LSTM layers traverse the sequence $\boldsymbol{e}_i$, producing sequences of vectors $\boldsymbol{h}_{if}$ and $\boldsymbol{h}_{ib}$ respectively, which are summed together in the coming sum layer.

Following the preprocessing method in (Lin et al., 2009), relations with too few instances (*Contingency.Condition, Pragmatic Condition; Comparison.Pragmatic Contrast, Pragmatic Concession; Expansion.Exception*) are removed during training and evaluation, resulting in 11 types of relations. Among instances annotated with two relation senses, we only use the first sense.

The model is implemented in Keras[6], which is capable of running on top of Theano. We use word embeddings of 300 dimensions, which are trained on the original English side of the parallel corpora as well as PDTB with the Skip-gram architecture in *Word2Vec* (Mikolov et al., 2013). We initialize the weights with uniform random; use standard cross-entropy as our loss function; employ Adagrad as the optimization algorithm of choice and set dropout layers after the embedding layer and output layer with a drop rate of 0.2 and 0.5 respectively. Each LSTM has a vector dimension of 300, matching the embedding size.

[6]https://keras.io/

Relation	intra-	inter-	Total
explicit → explicit	199,047	111,090	310,137
explicit → implicit	101,381	29,964	131,345
implicit → explicit	77,228	25,086	102,314

[1] "→" means from source to target side.

Table 1: Numbers of intra/inter-sentence samples extracted from parallel corpora.

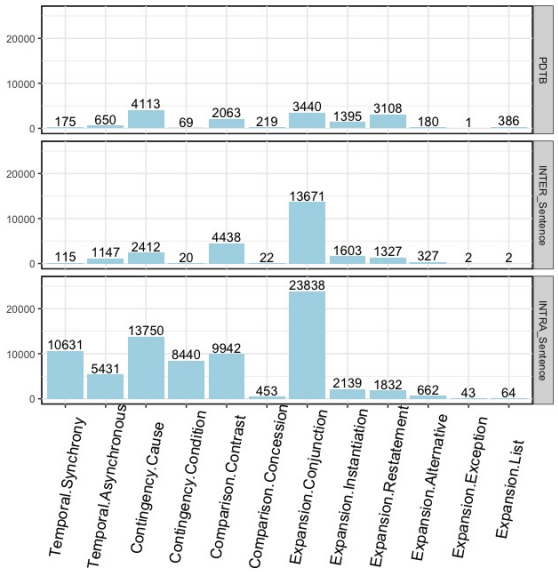

Figure 3: Relation sense distribution of implicit relations in PDTB and the extra intra- and inter-sentence samples

5 Distribution of additional instances

In total, 102,314 implicit discourse relation samples are extracted, of which 25,086 are inter-sentential relations and 77,228 are intra-sentential[7]. Inter-sentential relations are much less abundant because stricter screening strategy is applied (the end of point 3 in Section 3). From Table 1 we can also see that majority of DCs in the

[7]A dataset containing these additional instances will be made available to researchers upon publication of the paper.

Models		PDTB Test Set	Cross Validation
Most common class		25.36	25.59
Lin et al. (2009)		40.20	-
Qin et al. (2016)		43.81	-
Qin et al. (2017)		44.65	-
Rutherford et al. (2017)		39.56	-
Shi and Demberg (2017) (no surface features)		37.68	34.44
Ours	PDTB only	34.32	30.01
	PDTB + inter-sentential samples	42.29	34.14
	PDTB + intra-sentential samples	44.29	35.08
	PDTB + all samples	**45.50**	**37.84**

[1] "-" means no result currently.

Table 2: Accuracy of 11-way classification of implicit discourse relations on PDTB test set and by cross validation.

source side have been translated into the target side explicitly.

Figure 3 compares the distribution of relation senses among the annotated implicit relations in the PDTB and our extracted samples. The relation distribution generally corresponds to the distribution in PDTB, but some relations, such as *Temporal* and *Contingency.Condition*, are particularly numerous in the intra-sentential samples.

6 Results

We compare our model with current state-of-the-art models that were evaluated under the same setting (11-way classification, PDTB section 23 as test set) (Qin et al., 2016, 2017; Rutherford et al., 2017), as well as a model based on linguistic features (Lin et al., 2009) that uses this setting for evaluation.

Qin et al. (2017) developed an adversarial model, which consists of two CNNs in which arguments are represented separately, a four-layer Perceptron and a dense layer for classification, to enable an adaptive imitation scheme through competition between the implicit network and a rival feature discriminator. Our model substantially differs from that setup, as it uses a much simpler network architecture and represents the two discourse relation arguments jointly, i.e. without knowledge of the arguments' spans. We can see that our baseline model performs substantially less well than the state of the art, and also less well than (Shi and Demberg, 2017), who also use an LSTM but represent discourse relational arguments separately. As adding training data can be expected to be largely orthogonal to the choice of classification model, we are here most interested in seeing whether adding the new instances improves over the baseline model with identical architecture.

Table 2 shows that including the extra inter- and intra-sentential instances leads to very substantial improvements in classification accuracy. Using the additional data, our method not only improves performance by 11%-points on the PDTB test set compared to training on the PDTB implicit relations only, but also outperforms much more complex neural network models (Qin et al., 2016, 2017) on this task.

The evaluation using cross-validation (around 8% point improvement over the baseline) furthermore shows that the obtained improvements do not only hold for the PDTB standard test set but also are stable across the whole PDTB data. These results strongly support the effectiveness of the implicit relation samples mined from parallel texts.

The accuracies reported for our models are based on 10 repeat-runs with different initializations of the network. This allows us to show the amount of variance in results we obtained in Figure 4. We found that results sometimes varied a lot between different runs, and would therefore like to encourage others in the field to also report variability due to initialization or other random factors. For instance, our best run achieved 49.84% accuracy on the PDTB test set trained with all additional instances, while mean performance for that setting is 45.50% accuracy. Variances were substantially smaller for the cross-validation setting, as the number of overall instances going into the

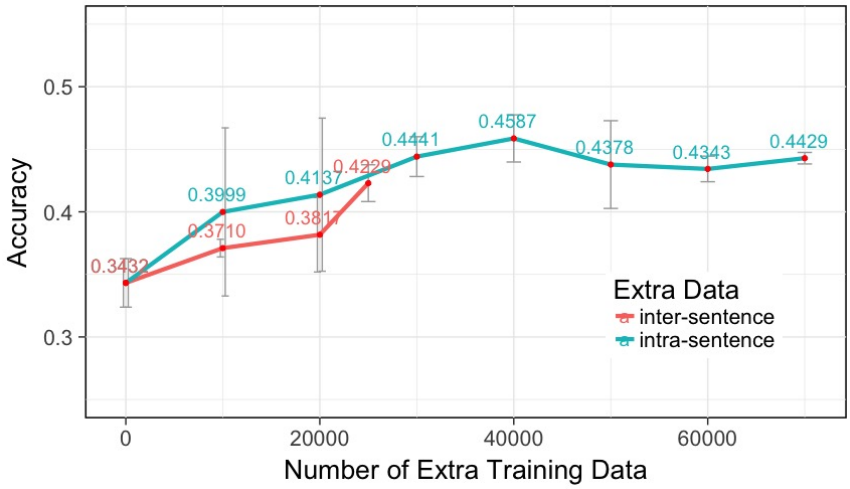

Figure 4: Average and variance of classification accuracy evaluated on the PDTB test set with different sample size.

evaluation is a lot larger in this setting, and hence yields more stable performance estimates.

6.1 Qualitative Analysis

In order to illustrate what kinds of instances our method extracts, we show an instances below. The underlined DC is the explicit DC identified in the back-translated target text; the discourse relation is automatically classified based on the back-translation.

3. *[Justice demands it.]$_{Arg1}$* ***but*** *[The minister refuses.]$_{Arg2}$*

   — Comparison.Contrast

One strength of the proposed method is that it can mine and label discourse relations that are not commonly regarded as discourse relations and hence not annotated in PDTB. Below are some examples where the bold DC was identified in the (back-)translation:

4. *A conservative member was kicked out of his caucus for defending Nova Scotians.*

   — **because**, Contingency.Cause

5. *A failure to do so would affect our attitude to their eventual accession.*

   —***if***, *Contingency.Condition*

These extra samples are in fact an invaluable resource of discourse-informative patterns, which are not available to discourse relation parsers that are trained only on the PDTB dataset. These cases provide evidence that our proposed method can not only provide instances that are similar to implicit labelled instances, but detect additional patterns, as attempted in (Mihăilă and Ananiadou, 2014; Hidey and McKeown, 2016) for causal relations, and generalize from the semantic content observed in such relations to actual implicit discourse relations.

For example, as reported in Section 5, numerous *Temporal* relations are mined from the parallel corpus. These include cases where the original text contained a verbal construction which expresses the temporal relation, which through back-translation gets expressed as a discourse relation, or where explicit relations include gerunds in the Arg2, e.g.

"any plan takes time to have the effect required"→ "before getting the effect required"

"how much longer do women have to wait for fairness?" → "before women have fairness."

"having gone over the estimates" → "after going over the estimates."

(source text followed by (back-)translation, where the explicitated DC is underlined).

In this work, we only extracted inter- and intra-sentential discourse relations, but the method can be in principle extended to other discourse relations that are not annotated in the PDTB, such as implicit relation between non-consecutive sentences. Discourse parsers that identify a larger range of relations are more useful in end applications. More importantly, identification of discourse-informative linguistic patterns by the proposed method opens the opportunity to mine extra samples under a monolingual setting and further improve classification performance.

6.2 Quantitative Analysis

In order to get detailed insights on how much extra data is most beneficial to the task, we also trained our classifier with different numbers of additional extracted samples. Figure 4 compares the classification accuracy when training on incremental number of extra instances. We find that the performance increases with samples size, but plateaus after $40,000$ intra-sentential samples.

In fact, this sample size produces the highest averaged classification accuracy of 45.87%, which is even higher than our model which includes all extracted samples. A possible reason for not seeing further improvement in adding more intra-sentential examples is the difference in distribution and properties of these extra samples compared to the PDTB data. We also experimented with training on the parallel-text samples only (i.e., without any PDTB training samples), but the result was worse than using PDTB only. Adding more *inter-sentential* samples might further improve the performance, as these instances are closer to the PDTB data.

6.3 Methodological Discussion

Our proposed method uses back-translated target DCs to label implicit relations. The quality of the relation label is intrinsically subject to the translation policy of the parallel corpora and also extrinsically subject to the accuracy of explicit DC classification by the end-to-end parser and the quality of the MT system. For example, a particularly high proportion of *Contingency.Condition* relations is found in the intra-sentential samples. Analyzing these samples, we found numerous instances where the word 'if' is wrongly identified as a DC (e.g. *He asked if it was correct.*). It is not surprising to have noisy samples extracted because limited screening strategy is applied in the current method.

As a reference for the quality of the relation label produced, we analysed the intra-sentential relations in the parallel corpus that are explicit on the source side and also in the (back-)translation. We found that 68% of the originally explicit DCs are (back-)translated to the same explicit DCs and 75% to DCs of the same level-2 sense, according to automatic explicit DC classification of the end-to-end parser.

7 Conclusion and Future work

We showed that explicitation during human translation can provide a valuable signal for expanding datasets for implicit discourse relations. As the expansion of training instances is orthogonal to the mechanism of DR classification, this method can be applied to improve any methods of implicit DR classification.

We see plenty of room for further improvement by controlling the sample quality, such as selection based on explicit discourse connective identification confidence, restraining the discourse relation structure, identifying Arg1 and Arg2 such that approaches which use two separate representations for arguments instead of a single concatenated vector become possible, reducing language-specific bias by mining from parallel corpora of other language pairs, and fine-tuning the MT system for discourse connective translation. We leave the exploration of these areas to future work.

Acknowledgments

We would like to thank the anonymous reviewers for their careful reading, valuable and insightful comments. This work was funded by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding".

References

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, pages 2669–2684.

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sofia, Bulgaria, pages 69–73.

Bruno Cartoni, Sandrine Zufferey, and Thomas Meyer. 2013. Annotating the meaning of discourse connectives by looking at their translation: The translation-spotting technique. *D&D* 4(2):65–86.

Bruno Cartoni, Sandrine Zufferey, Thomas Meyer, and Andrei Popescu-Belis. 2011. How comparable are parallel corpora? measuring the distribution of general vocabulary and connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*. Association for Computational Linguistics, pages 78–86.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1726–1735.

Philipp Cimiano, Uwe Reyle, and Jasmin Šarić. 2005. Ontology-driven discourse analysis for information extraction. *Data & Knowledge Engineering* 55(1):59–83.

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*. Association for Computational Linguistics, pages 1602–1613.

Francisco Guzmán, Shafiq R. 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*. Association for Computational Linguistics, pages 687–698.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 399–409.

Christopher Hidey and Kathleen McKeown. 2016. Identifying causal relation using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1424–1433.

Jet Hoek and Sandrine Zufferey. 2015. Factors influencing the implicitation of discourse relations across languages. In *Proceedings 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-11)*. TiCC, Tilburg center for Cognition and Communication, pages 39–45.

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*. Association for Computational Linguistics, 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*. Association for Computational Linguistics, pages 332–342.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810* .

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1318–1326.

Majid Laali and Leila Kosseim. 2014. Inducing discourse connectives from parallel texts. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING-2014)*. pages 610–619.

Man Lan, Yu Xu, Zheng-Yu Niu, et al. 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 476–485.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin R. Batchelor, and Dietrich Rebholz-Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 747–757.

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*. Association for Computational Linguistics, Singapore, pages 343–351.

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.

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

Sameer Maskey and Julia Hirschberg. 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Ninth European Conference on Speech Communication and Technology*.

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

Claudiu Mihăilă and Sophia Ananiadou. 2014. Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomedical engineering online* 13(2):1.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

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*. Association for Computational Linguistics, pages 108–112.

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*. Association for Computational Linguistics, Suntec, Singapore, pages 683–691.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K. Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*. Manchester, UK, pages 85–88.

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*. European Language Resources Association, Marrakech, Morocco.

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)*. Association for Computational Linguistics, Vancouver, Canada, pages 1006–1017.

Ella Rabinovich, Shuly Wintner, and Ofek Luis Lewinsohn. 2015. The haifa corpus of translationese. *arXiv preprint arXiv:1509.03611* .

Samuel Rönnqvist, Niko Schenk, and Christian Chiarcos. 2017. A recurrent neural model with attention for the recognition of chinese implicit discourse relations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 256–262.

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*. Association for Computational Linguistics, 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*. Association for Computational Linguistics, pages 645–654.

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*. Association for Computational Linguistics, pages 799–808.

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*. Association for Computational Linguistics, pages 150–156.

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

Yannick Versley. 2010. Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In *AEPC*. pages 83–82.

Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2015. Word embedding for recurrent neural network based tts synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, pages 4879–4883.

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.

Changxing Wu, Xiaodong Shi, Yidong Chen, Yanzhou Huang, and Jinsong Su. 2016. Bilingually-constrained synthetic data for implicit discourse relation recognition. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2306–2312.

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*. Association for Computational Linguistics, pages 1–16.

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*. Association for Computational Linguistics, pages 1–19.

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*. Association for Computational Linguistics, pages 1834–1839.

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*. Association for Computational Linguistics, pages 2230–2235.

Lanjun Zhou, Wei Gao, Bin Li, Zhong Wei, and Kam-Fai Wong. 2012. Cross-lingual identification of ambiguous discourse connectives for resource-poor language. In *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee, pages 1409–1418.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 207–212.

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 (COLING 2010)*. Association for Computational Linguistics, Beijing, China, pages 1507–1514.

Sandrine Zufferey. 2016. Discourse connectives across languages. *Languages in Contrast* 16(2):264–279.