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

Abstract

Implicit discourse relation classification is one of the most challenging and important tasks in discourse parsing, due to the lack of connective as strong linguistic cues. A principle bottleneck to further improvement is the shortage of training data (ca.~16k instances in the 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.

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 BIBREF1 , BIBREF2 , question answering BIBREF3 , BIBREF4 , machine translation BIBREF5 , BIBREF6 , and so on.

Discourse relations can be marked explicitly using a discourse connective or discourse adverbial such as “because”, “but”, “however”, see example SECREF1 . Explicitly marked relations are relatively easy to classify automatically BIBREF7 . In example SECREF2 , 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 BIBREF8 , but is much harder to classify automatically.

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 BIBREF9 , making human-annotated relations the currently only source for training relation classification.

Due to the limited size of available training data, several approaches have been proposed for acquiring additional training data using automatic methods BIBREF10 , BIBREF11 . The most promising approach so far, BIBREF0 , 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, BIBREF0 showed that such instances can be used for acquiring additional labeled text.

BIBREF0 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.

Related Work

Recent methods for discourse relation classification have increasingly relied on neural network architectures. However, with the high number of parameters to be trained in more and more complicated deep neural network architectures, the demand of more reliable annotated data has become even more urgent. Data extension has been a longstanding goal in implicit discourse classification. BIBREF10 proposed to differentiate typical and atypical examples for each relation and augment training data for implicit only by typical explicits. BIBREF11 designed criteria for selecting explicit samples in which connectives can be omitted without changing the interpretation of the discourse. More recently, BIBREF0 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 extended theirs by employing document-aligned cross-lingual parallel corpora and majority votes to get more reliable and in-topic annotated implicit discourse relation instances.

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 BIBREF0 , as illustrated in Figure FIGREF3 , with the following differences: First, 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. Secondly, we use a statistical machine translation system instead of a neural one for more stable translations.

Machine Translation

We train three MT systems to back-translate French, German and Czech to English. To have words alignments, better and stable back-translations, we employ a statistical machine translation system Moses BIBREF12 , 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).

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 which could be labeled with high confidence.

Data

Europarl Corpora The parallel corpora used here are from Europarl BIBREF13 , 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) It is the largest 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.

Implicit discourse relation classification

To evaluate whether the extracted data is helpful to this task, we use a simple and effective bidirectional Long Short-Term Memory (LSTM) network. 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 inputed into a softmax layer for prediction.

Implementation: The model is implemented in Pytorch. All the parameters are initialized with uniform 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 ourput layer with drop rates of 0.5 and 0.2 respectively. The word vectors are pre-trained word embedding from Word2Vec.

Settings: We follow the previous works and evaluate our data on second-level 11-ways classification on PDTB with 3 settings: BIBREF14 (denotes as PDTB-Lin) uses sections 2-21, 22 and 23 as train, dev and test set; BIBREF15 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 BIBREF16 . For each experiment, the additional data is only added into the training set.

Results

Figure FIGREF11 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. The only exceptions are Expansion.Conjunction relations from the German translations, which are over-represented, and Expansion.Restatement relations which are under-represented based on our back-translation method.

Figure FIGREF14 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.

Table TABREF7 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, introducing noise into the system. The size of the extra data used in BIBREF0 is about 10 times larger than our 2-votes data, as they relied on additional training data (which we could not use in this experiment, as there is no pairing with translations into other languages) and exploited also intra-sentential instances. While we don't match the performance of BIBREF0 on the PDTB-Lin test set, the high quality translation data shows better generalisability by outperforming all other settings in the cross-validation (which is based on 16 test instances, while the PDTB-Lin test set contains less than 800 instances and hence exhibits more variability in general).

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 recognized as a discourse relation marker by the automatic discourse parser:

Original English: I presided over a region crossed by heavy traffic from all over Europe...what is more, in 2002, two Member States of the European Union appealed to the European Court of Justice...

French: moreover (Expansion.Conjunction)

German: moreover (Expansion.Conjunction)

Czech: therefore (Contingency.Cause) after all

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.)

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: ...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: also (Expansion.Conjunction)

German: therefore (Contingency.Cause)

Czech: therefore (Contingency.Cause)

Case 3: Discrepancies between connectives in back-translation can also be due to differences in how translators interpreted the original text:

Original English: ...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...

French: however (Comparison.Contrast)

German: therefore (Contingency.Cause)

Czech: in addition (Expansion.Conjunction)

Case 4: Implicit relations can co-occur with marked discourse relations BIBREF17 , 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...

French: but (Comparison.Contrast) there is more

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. A qualitative analysis furthermore shows that the strength of the method partially stems from being able to learn additional discourse cues which are typically translated consistently, and suggests that our method may also be used for identifying multiple relations holding between two arguments.