Neurocomputing 000 (2017) 1–11

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Leveraging bilingually-constrained synthetic data via multi-task neural networks for implicit discourse relation recognition

Changxing Wu[a,b], Xiaodong Shi[a,b,*], Yidong Chen[a,b], Yanzhou Huang[a,b], Jinsong Su[c]

[a] Fujian Key Lab of the Brain-like Intelligent Systems, Xiamen University, Xiamen 361005, Fujian, China
[b] Department of Cognitive Science, School of Information Science and Technology, Xiamen University, Xiamen 361005, Fujian, China
[c] School of Software, Xiamen University, Xiamen 361005, Fujian, China

ARTICLE INFO

Article history:
Received 22 September 2016
Revised 7 January 2017
Accepted 27 February 2017
Available online xxx

Communicated by Dr. Y. Chang

Keywords:
Bilingually-constrained synthetic implicit data
Multi-task learning
Implicit discourse relation recognition
Neural network

ABSTRACT

Recognizing implicit discourse relations is an important but challenging task in discourse understanding. To alleviate the shortage of labeled data, previous work automatically generates synthetic implicit data (*SynData*) as additional training data, by removing connectives from explicit discourse instances. Although *SynData* has been proven useful for implicit discourse relation recognition, it also has the meaning shift problem and the domain problem. In this paper, we first propose to use bilingually-constrained synthetic implicit data (*BiSynData*) to enrich the training data, which can alleviate the drawbacks of *SynData*. Our *BiSynData* is constructed from a bilingual sentence-aligned corpus according to the implicit/explicit mismatch between different languages. Then we design a multi-task neural network model to incorporate our *BiSynData* to benefit implicit discourse relation recognition. Experimental results on both the English PDTB and Chinese CDTB data sets show that our proposed method achieves significant improvements over baselines using *SynData*.

1. Introduction

Discovering the discourse relation between two sentences/clauses is crucial to understanding the meaning of a coherent text, and also beneficial to many downstream natural language processing (NLP) applications, such as question answering and machine translation. A discourse relation instance is usually defined as a connective (denoted as *Conn*, e.g. *but*) taking two arguments (as *Arg*1 and *Arg*2 respectively), where the arguments can be sentences or clauses. According to the existence of connectives or not, discourse relation instances are grouped into two categories: explicit instances, denoted as (*Arg*1, *Arg*2, *Conn*), and implicit instances as (*Arg*1, *Arg*2). Explicit discourse relations can be easily distinguished, with an accuracy of about 94% [28]. By contrast, implicit discourse relation recognition (DRR_{imp}) remains a challenging task due to the absence of strong surface clues like discourse connectives. Therefore, most work resorts to large amounts of manually designed features [3,16,20,27,31,32], or distributed features automatically learned via neural network models [4,10,37]. The above methods usually suffer from limited labeled data.

To address the shortage of labeled data, [23] automatically generate training data from explicit instances using a pattern-based approach. They remove discourse connectives from extracted explicit instances and map them into discourse relations. For example, explicit instances signaled by *but* can be potentially used as additional training data for the *Comparison* relation in implicit discourse relation recognition. These data are usually called *synthetic implicit data* (*SynData*). However, [33] argue that *SynData* has two drawbacks: 1) meaning shifts in some cases when removing discourse connectives[1], and 2) a different word distribution with the *real implicit data*, in other words, the explicit/implicit domain problem. They also show that using *SynData* directly degrades the performance of DRR_{imp}. Recent work seeks to learn valuable information from *SynData* while filtering noise, via domain adaptation [2], classifying connectives [30] or multi-task learning [19], and shows promising results.

Different from previous work, we propose to construct *bilingually-constrained synthetic implicit data* (called *BiSynData*) for implicit discourse relation recognition, which can alleviate the drawbacks of *SynData*. Our method is inspired by the findings that a discourse instance expressed implicitly in one language may be

* Corresponding author at: Fujian Key Lab of the Brain-like Intelligent Systems, Xiamen University, Xiamen 361005, Fujian, China.
E-mail addresses: mandel@xmu.edu.cn, mandelshi@gmail.com (X. Shi).

[1] Considering the explicit instance: *I am eager to go home for the vacation. Nonetheless, I will book a flight to Beijing.*, one would infer a *Contingency* relation rather than the *Comparison* relation if *nonetheless* is dropped.

http://dx.doi.org/10.1016/j.neucom.2017.02.084
0925-2312/

ch: [社会 认为 有 青少年 问题,]$_{Arg1}$
society reckon existence youth problems,
implicit=但是 [很多 青少年 认为 自己 没 问题.]$_{Arg2}$
but many young people think themselves no problems.
en: [Society reckons the existence of youth problems,]$_{Arg1}$
but [many young people do not think there is anything wrong with them.]$_{Arg2}$

Fig. 1. An example illustrating the implicit/explicit mismatch between Chinese (*ch*) and English (*en*). A Chinese implicit instance is translated into an English explicit one with *but*.

expressed explicitly in another. For example, [38] show that the connectives in Chinese omit much more frequently than those in English with about 82.0% vs. 54.5%. [14] further argue that there are about 23.3% implicit/explicit mismatches between Chinese/English instances. As illustrated in Fig. 1, a Chinese implicit instance where the connective 但是 is absent, is translated into an English explicit one with the connective *but*. Intuitively, the Chinese instance is a *real* implicit one which can be signaled by the English connective *but*. Hence, it could potentially serve as additional training data for the Chinese DRR_{imp}, avoiding the domain problem of *SynData*. Meanwhile, for the English explicit instance, it is very likely that removing *but* would not lose any information since its Chinese counterpart 但是 can be omitted when translation. Therefore it could be used for the English DRR_{imp}, alleviating the meaning shift problem of *SynData*. Based on the above analysis, we believe *BiSynData* is more suitable for implicit discourse relation recognition than *SynData*.

We incorporate *BiSynData* via multi-task learning to improve the performance of DRR_{imp}. Specifically, we design a simple multi-task neural network model which synthesizes DRR_{imp} (as the main task) and connective classification tasks on *BiSynData* (as the auxiliary task). For each argument in an implicit instance, we simply average embeddings of words to represent it. Then we use multiple non-linear hidden layers to catch the complicated interactions between two arguments. To combine the main and auxiliary tasks, we investigate two parameter sharing strategies to learn the connections between them. With our multi-task neural network, the main and auxiliary tasks are trained simultaneously and learn from each other through their connections.

Our method achieves significant improvements over baselines using *SynData*, on both the English PDTB [29] and Chinese CDTB [15] data sets. Moreover, on the PDTB, our method performs better than the strong baseline [19] which uses additional labeled data. The major contribution of this paper is that we introduce *BiSynData* to implicit discourse relation recognition for the first time. We also develop a simple and effective multi-task neural network model to incorporate *BiSynData*.

The contents of the paper are organized as follows. We extract our *BiSynData* from a Chinese-English sentence-aligned corpus in Section 2, and introduce our multi-task neural network model in Section 3. Then we conduct experiments to validate the effectiveness of our proposed method in Section 4. Finally, we review the related work in Section 5 and draw conclusions in Section 6.

2. *BiSynData*

Formally, given a Chinese–English sentence pair (S_{ch}, S_{en}), we try to find an English explicit instance ($Arg1_{en}$, $Arg2_{en}$, $Conn_{en}$) in S_{en}, and a Chinese implicit instance ($Arg1_{ch}$, $Arg2_{ch}$) in S_{ch}, where ($Arg1_{en}$, $Arg2_{en}$, $Conn_{en}$) is the translation of ($Arg1_{ch}$, $Arg2_{ch}$). In most cases discourse relations should be preserved during translation, so the connective $Conn_{en}$ is potentially a strong indicator of the discourse relation between not only $Arg1_{en}$ and $Arg2_{en}$ but also $Arg1_{ch}$ and $Arg2_{ch}$. Therefore, we can construct two synthetic implicit instances labeled by $Conn_{en}$, denoted as $\langle(Arg1_{en}, Arg2_{en}), Conn_{en}\rangle$ and $\langle(Arg1_{ch}, Arg2_{ch}), Conn_{en}\rangle$, respectively. Similarly, considering a sentence pair with a Chinese explicit instance ($Arg1_{ch}$, $Arg2_{ch}$, $Conn_{ch}$) and an English implicit one ($Arg1_{en}$, $Arg2_{en}$), the Chinese connective $Conn_{ch}$ also indicates the discourse relation. In this case, we can obtain two synthetic implicit instances labeled by $Conn_{ch}$: as $\langle(Arg1_{en}, Arg2_{en}), Conn_{ch}\rangle$ and $\langle(Arg1_{ch}, Arg2_{ch}), Conn_{ch}\rangle$, respectively. We refer to the above four kinds of synthetic instances as *BiSynData* because they are constructed according to the bilingual implicit/explicit mismatch.

Algorithm 1 Extracting *BiSynData*.

Input: C is a Chinese/English sentence-aligned corpus, A_w is the word alignments between bilingual sentence pairs.
Output: *BiSynData*

```
1: BiSynData ← ϕ
2: for each sentence pair (S_ch, S_en) in C do
3:   if exist a Chinese connective in S_ch then
4:     continue
5:   end if
6:   if exist an explicit instance (Arg1_en, Arg2_en, Conn_en) in S_en
     then
7:     Find its translation (Arg1_ch, Arg2_ch) in S_ch according to A_w.
8:     Add ⟨(Arg1_en, Arg2_en), Conn_en⟩ into BiSynData.
9:     Add ⟨(Arg1_ch, Arg2_ch), Conn_en⟩ into BiSynData.
10:  end if
11: end for
12: for each sentence pair (S_ch, S_en) in C do
13:  if exist an English connective in S_en then
14:    continue
15:  end if
16:  if exist an explicit instance (Arg1_ch, Arg2_ch, Conn_ch) in S_ch
     then
17:    Find its translation (Arg1_en, Arg2_en) in S_en according to A_w.
18:    Add ⟨(Arg1_en, Arg2_en), Conn_ch⟩ into BiSynData.
19:    Add ⟨(Arg1_ch, Arg2_ch), Conn_ch⟩ into BiSynData.
20:  end if
21: end for
22: return BiSynData
```

The procedure of extracting *BiSynData* is described in Algorithm 1. First, we simply discard a sentence pair (S_{ch}, S_{en}) if a Chinese connective exists in S_{ch} (Line 3–5). This releases us from the need to distinguish whether the connective reflects a discourse relation[2]. Then we use the pdtb-parser toolkit [17] to identify whether there is an English explicit instance in S_{en} (Line 6). Finally, given positions of the English explicit instance and word alignments between S_{ch} and S_{en}, we obtain a Chinese implicit instance which is the translation of the English explicit

[2] For example, the Chinese word 和 (*and*) can either function as a discourse connective to join two *Expansion* events, or be just used to link two *nouns* in a phrase.

Please cite this article as: C. Wu et al., Leveraging bilingually-constrained synthetic data via multi-task neural networks for implicit discourse relation recognition, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.02.084