A Knowledge-Augmented Neural Network Model for Implicit Discourse Relation Classification

Yudai Kishimoto **Yugo Murawaki** **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{kishimoto, murawaki, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Identifying discourse relations that are not overtly marked with discourse connectives remains a challenging problem. The absence of explicit clues indicates a need for the combination of world knowledge and weak contextual clues, which can hardly be learned from a small amount of manually annotated data. In this paper, we address this problem by augmenting the input text with external knowledge and context and by adopting a neural network model that can effectively handle the augmented text. Experiments show that external knowledge did improve the classification accuracy. On the other hand, contextual information provided no significant gain for implicit discourse relations while it worked for explicit ones.

1 Introduction

Discourse relation recognition has a wide variety of potential applications including summarization (Louis et al., 2010), sentiment analysis (Somasundaran et al., 2009) and machine translation (Meyer et al., 2015). In one of the two most prevalent discourse treebanks, the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008), discourse relations are conventionally divided into two types: explicit and implicit. Explicit relations are overtly marked with discourse connectives such as "because" and "however." Because of these strong cues, explicit relations are relatively easy to classify (Pitler et al., 2008; Xue et al., 2016). By contrast, implicit relations lack discourse connectives and classifying such relations remains a challenging problem.

Recent studies on implicit discourse relation classification have shown success in applying various neural network models including feedforward networks (Zhang et al., 2015; Schenk et al., 2016), convolutional neural networks (Mihaylov and Frank, 2016; Wang and Lan, 2016) and bidirectional LSTM (bi-LSTM) (Chen et al., 2016; Liu and Li, 2016; Dai and Huang, 2018). Although these studies on network engineering report performance improvement, Rutherford et al. (2017) demonstrated that a simple feed-forward neural network was astonishingly competitive, outperforming LSTM- and Tree LSTM-based models. He claimed that training data for implicit discourse relation classification were too small to train powerful neural networks like LSTM. This motivates us to view this task from a different perspective.

We argue that the neural network models need to be provided with world knowledge, which can hardly be learned from a small amount of manually annotated data. To see this, suppose that we want to classify the discourse relation of the following pair of text spans (referred to as `Arg1` (in *italic*) and `Arg2` (in **bold**) throughout this paper):

> `Arg1`: *Not counting the extraordinary charge it would have had a net loss of $3.1 million, or seven cents a share*
> `Arg2`: **A year earlier, it had profit of $7.5 million, or 18 cents a share**
> (Comparison.Contrast)

We can easily recognize that the discourse relation between this pair is Comparison.Contrast partly because we know that "loss" in `Arg1` and "profit" in `Arg2` are antonyms. However, it is difficult for a

Proceedings of the 27th International Conference on Computational Linguistics, pages 584–595
Santa Fe, New Mexico, USA, August 20-26, 2018.

neural network model trained on small training data to recognize the antonymy. Although many recent studies make use of word2vec word embeddings (Mikolov et al., 2013), which are typically trained with a large amount of unannotated data, it is well known that synonyms and antonyms are distributionally similar and thus are hardly distinguishable (Ono et al., 2015). For this reason, we need to look for different knowledge sources as well as an efficient way to integrate them into neural network models.

In this paper, we use MAGE-GRU (Dhingra et al., 2017) to encode external knowledge. It is a straightforward extension to the Gated Recurrent Unit (GRU) (Cho et al., 2014). The input to MAGE-GRU is no longer a sequence of words but a directed acyclic graph in which the word sequence is augmented with edges between arbitrarily distant words for which external knowledge suggests some explicit signals such as antonymy. Thus "loss" and "profit" in the example above are directly connected, making a downstream network layer more easily classify the discourse relation.

In the experiments, we augmented the inputs with ConceptNet (Speer and Havasi, 2012) and coreference resolution. We found that MAGE-GRU significantly outperformed others when ConceptNet was used.

While recurrent neural networks have considerable difficulty in capturing long-range dependencies, MAGE-GRU is expected to mitigate this problem because it creates shortcuts within word sequences. This leads us to explore another question: Do `Arg1` and `Arg2` provide sufficient information to determine discourse relations?

Let us consider the following example:

> Good service programs require recruitment, screening, training and supervision – all of high quality. They involve stipends to participants. Full-time residential programs also require housing and full-time supervision; *they are particularly expensive – more per participant than a year at Stanford or Yale*. **Non-residential programs are cheaper, but good ones still come to some $10,000 a year** (Comparison.Contrast)

In this example, "Non-residential programs" in `Arg2` is contrasted with "they" in `Arg1`. To recognize the fact, we need to resolve the pronoun "they" by looking back at the text preceding `Arg1` to find the antecedent "Full-time residential programs." Although other clues such as the pair of "expensive" and "cheaper" are present, contextual information makes it easier to classify this argument pair as Comparison.Contrast. Although the current standard approach to this task is to use the pair of `Arg1` and `Arg2` out of context, the computer might also benefit from the wider context, given the power of MAGE-GRU.

We compared the performance of MAGE-GRU with and without the text chunk that preceded a given argument pair in the paragraph. It turned out that contextual information provided no significant improvement for implicit discourse relations. However, we also found that contextual information yielded a significant gain for explicit discourse relations. The results appear to strengthen the observation that explicit and implicit discourse relations are dissimilar (Prasad et al., 2014).

2 Related Work

Before neural networks were introduced to the task of implicit discourse relation classification, Lin et al. (2009) proposed a linear classifier that was based on various lexical and syntactic features and was combined with extensive feature selection. Rutherford et al. (2017) built a simple feedforward neural network model where only one pooling layer and one hidden layer were stacked on top of word embeddings. They reported that the simple model outperformed LSTM- and Tree LSTM-based models but lost to Lin et al. (2009).

LSTM has demonstrated success in a wide range of NLP tasks including implicit discourse relation classification. Chen et al. (2016) proposed a combination of a bi-LSTM and a gated relevance network while Liu and Li (2016) combined a bi-LSTM with multi-level attentions. Dai and Huang (2018) proposed a combination of a word-level bi-LSTM and a paragraph-level neural networks. Rutherford et al. (2017) argued that the annotated corpus was too small to train LSTM-based models, however. Note that the present work is complementary to these studies. Since our model is a straightforward extension to

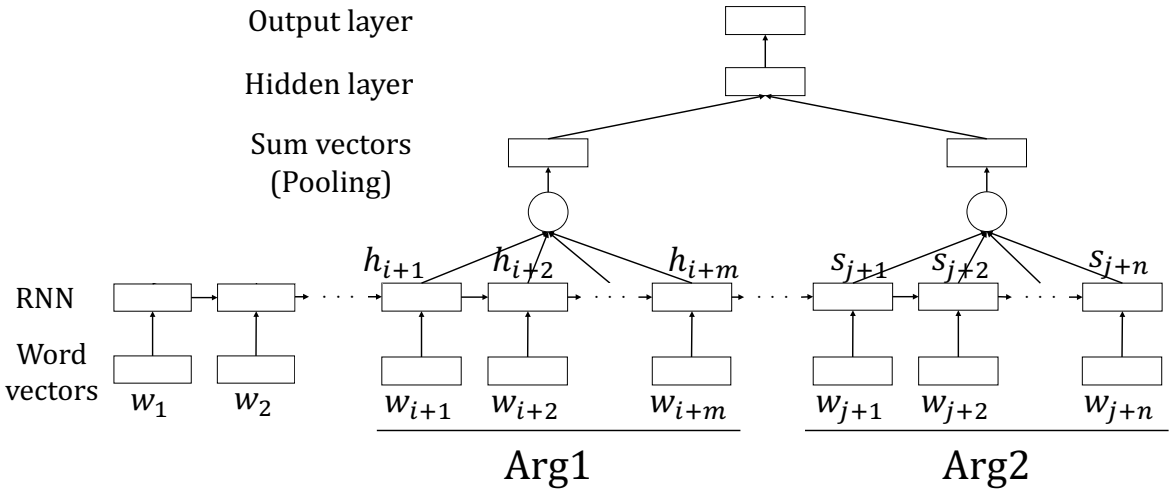

Figure 1: System architecture.

the recurrent neural network, it can easily be combined with other neural network models that are built on top of LSTMs.

Another popular choice of network is convolutional neural networks (CNN). Wang and Lan (2016) proposed an end-to-end shallow discourse parser. In their pipeline system, a CNN is used for non-explicit relation classification.[1] It ranked second on the blind datasets in the CoNLL 2016 Shared Task. Qin et al. (2016) proposed a combination of a bi-LSTM and CNNs. They constructed character-based word representations by transforming character embeddings with CNN and bi-LSTM layers. Another CNN layer was used to extract an argument representation from a sequence of words.

Some recent studies exploited external knowledge sources and some of them were reported to improve the performance of implicit discourse relation classification (Rutherford and Xue, 2014). In the closed track of the CoNLL 2016 Shared Task (Xue et al., 2016), the organizers allowed participants to use a limited set of linguistic resources: Brown Clusters, VerbNet, a sentiment lexicon and an off-the-shelf word2vec model. In addition to these, Inquirer Tags, Levin classes and Modality were tested by Shi and Demberg (2017). They extracted features from these resources and added them to a neural network layer just before the output. Dhingra et al. (2017), who proposed MAGE-GRU, tested a more direct approach as a baseline, in which they appended to word embeddings a sequence of features which are fired by external knowledge. They found that MAGE-GRU consistently outperformed the baseline in various tasks when coreference resolution is used as external knowledge.

A huge performance gap between explicit and implicit relations leads some to transform explicit relations in unannotated corpora to generate pseudo-training data of implicit relations. Sporleder and Lascarides (2008) reported negative results, suggesting that explicit and implicit discourse relations were linguistically dissimilar. Rutherford and Xue (2015) worked on selecting discourse connectives that can be safely dropped. Qin et al. (2017) generated a different kind of pseudo-training data. They inserted to implicit relations *implicit connectives* PDTB annotators assigned to them. They used domain adversarial training to transfer knowledge from the recognition model supplied with implicit connectives to the model without connectives. These methods can be combined with our approach.

3 Proposed Method

The overall system architecture is shown in Figure 1. It is based on the RNN architecture of Rutherford et al. (2017) although we made two major modifications to it, which will be described in Sections 3.1 and 3.3. For each of `Arg1` and `Arg2`, a sequence of words are transformed into a sequence of hidden

[1]Non-explicit relation classification is closely related to implicit discourse relation classification but differs in the treatment of another type of relation, AltLex.

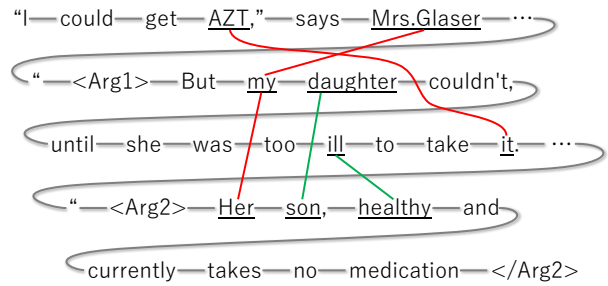

Figure 2: Sequence augmented with extra edges. Words in the sequence are connected by gray edges. In addition, pairs of words are connected by red edges if they are coreferential. Similarly, green edges denote antonymy.

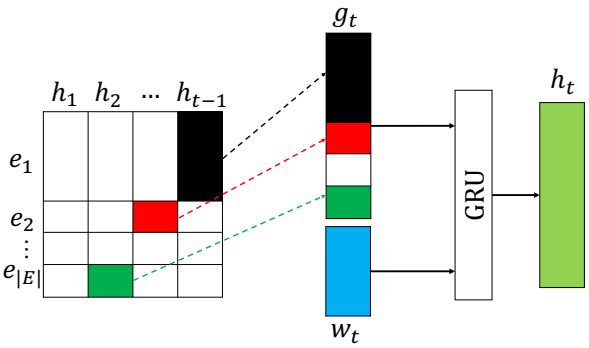

Figure 3: One step of MAGE-GRU.

vectors by an RNN. It is followed by a pooling operation, the concatenation of the two arguments, a linear transformation with non-linear activation, and a softmax classification to predict the relation. The linear transformation is designed to capture the interaction between the arguments.

3.1 MAGE-GRU

One major difference from the architecture of Rutherford et al. (2017) is that whereas Rutherford et al. (2017) used LSTM as the RNN component, we adopt MAGE-GRU (Dhingra et al., 2017). It extends the Gated Recurrent Unit (GRU) (Cho et al., 2014) such that it can incorporate external linguistic knowledge as explicit signals.

The input to MAGE-GRU is no longer a sequence of words but a directed acyclic graph as shown in Figure 2. The graph is constructed by augmenting the word sequence with edges between arbitrarily distant words. Sequential transitions are now treated as a special type of edges. An edge is added if external knowledge suggests an explicit signal for the given word pair. We represent a signal as a triplet (*word A*, *relation type*, *word B*). For example, green edges in Figure 2 denote antonymy: (daughter, `Antonym`, son) and (ill, `Antonym`, healthy). Edges are directed and we assume that *word A* always comes before *word B*. If relation types in external knowledge are asymmetric (e.g., (thinking, `Causes`, acting)), we create reverse relation types to keep the directionality. If external knowledge deals with a phrase, we use the last word for the triplet. We also constrain every node to have at most one incoming edge per relation type because this drastically simplifies computation. If *word B* has two or more candidates for *word A*, the nearest one is chosen.

Given a directed acyclic graph, MAGE-GRU outputs a hidden vector h_t at each time step t. One step of MAGE-GRU is illustrated in Figure 3. The peculiarity of MAGE-GRU is that whereas vanilla GRU receives word vector w_t and hidden vector h_{t-1} as the inputs, MAGE-GRU substitutes h_{t-1} with a special vector g_t. g_t is created by partially and selectively combining the history of hidden vectors, $h_1, \cdots, h_{t-1}$:

$$
\begin{aligned}
g_t &= [g_t^{e_1}; g_t^{e_2}; \cdots; g_t^{e_{|E|}}] \\
g_t^{e_i} &= \begin{cases} h_{t'}^{e_i} & \text{if } (x_{t'}, e_i, x_t) \text{ exists} \\ \mathbf{0} & \text{otherwise} \end{cases}
\end{aligned}
$$

where ; denotes vector concatenation. $h_{t'}$ is decomposed into $[h_{t'}^{e_1}; h_{t'}^{e_2}; \cdots; h_{t'}^{e_{|E|}}]$, where e_i $(1 \leq i \leq |E|)$ is a relation type. Thus $h_{t'}^{e_i}$ is a subvector of $h_{t'}$ corresponding to relation type e_i. In Figure 2, for example, when w_t is "healthy", g_t consists of $h_{\text{son}}^{\text{sequence}}$ due to the gray edge and $h_{\text{ill}}^{\text{antonym}}$ due to the green edge and zero vectors for the rest. Note that e_1 is reserved for ordinary sequential transitions.

3.2 External Knowledge

We use ConceptNet (Speer and Havasi, 2012) and coreference resolution as external knowledge. ConceptNet is an open-source project for building in a large linguistic knowledge base. It was constructed through various means including handcrafting (Open Mind Common Sense), automatic extraction from web pages such as Wikipedia, and gamification. ConceptNet covers not only lexical definitions but also common sense. Some relations in ConceptNet appear to be particularly useful for implicit discourse relation classification. For example, the relation `Causes` is closely related to the discourse relation Contingency.Cause. We expect the system to classify such discourse relations more easily when the input is augmented with ConceptNet.

We also use coreference resolution as external knowledge. As we discussed in Section 1, coreference resolution is a straightforward approach to connecting arguments to contextual information. Coreference resolution was also used by Dhingra et al. (2017) in their experiments.

3.3 Input Formats

Another major modification we make to the architecture of Rutherford et al. (2017) is about the formats of input sequences. While Rutherford et al. (2017) among others treated `Arg1` and `Arg2` separately, we combine them into a single sequence. Additionally, we append to the sequence a text chunk that precedes the arguments in the paragraph.[2]

We insert special tags, <Arg1>, </Arg1>, <Arg2>, and </Arg2>, into a given sequence to indicate where the arguments start and end. A similar technique is used in neural machine translation (Johnson et al., 2016). Although the PDTB annotates implicit relations on adjacent sentences, an argument pair does not necessarily form a complete span mainly because sub-sentential spans are sometimes selected as arguments (The PDTB Research Group, 2007). In addition, arguments sometimes contain non-argument text spans as in the following example (underlined):

> *At Quantum*, which is based in New York, *the trouble is magnified by the company's heavy dependence on plastics*. Once known as National Distillers & Chemical Corp., **the company exited the wine and spirits business and plowed more of its resources into plastics after Mr. Stookey took the chief executive's job in 1986**.
> (Contingency.Cause)

To cope with this problem, we enclose a non-argument span with <Skip> and </Skip>.

4 Experiments

4.1 Setup

4.1.1 Penn Discourse TreeBank

We evaluated our model's performance on the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008). It is the most popular and largest corpus of discourse relations in English. The annotation is done as another layer on Wall Street Journal sections of the Penn Treebank. Each discourse relation consists of two text spans (arguments) and a relation label. Arguments are annotated such that they are minimally required to infer the discourse relation.

Relation labels are organized as a 3-level hierarchy in the PDTB. However, it is too difficult for current systems to perform classification in its original form, and previous studies have used more coarse-grained relation labels. Popular settings include top-level one-versus-all binary classification (Pitler et al., 2009), top-level 4-way classification (Pitler et al., 2009; Ji and Eisenstein, 2015), second-level 11-way classification (Lin et al., 2009; Rutherford et al., 2017), and modified second-level classification for the CoNLL 2015 Shared Task (Xue et al., 2015). We used second-level 11-way classification in the experiments.

Following Shi and Demberg (2017), we conducted 10-fold cross validation using the whole corpus of sections 0–24 (referred to as Cross Validation). The standard approach (referred to as Most-used Split)

[2] Rönnqvist et al. (2017) also concatenated `Arg1` and `Arg2` in the task of Chinese implicit discourse relation classification. They did not incorporate contextual information, however.

Sense	Train	Dev	Test
Comparison.Concession	179	19	21
Comparison.Contrast	1,672	185	206
Contingency.Cause	3,332	370	411
Contingency.Pragmatic cause	57	6	6
Expansion.Alternative	146	16	18
Expansion.Conjunction	2,787	309	344
Expansion.Instantiation	1,131	125	139
Expansion.List	314	34	38
Expansion.Restatement	2,519	279	310
Temporal.Asynchronous	527	58	65
Temporal.Synchrony	143	15	17
Total	12,807	1,416	1,575

Table 1: The distribution of relation labels in the Cross Validation dataset.

Sense	Train	Dev	Test
Comparison.Concession	192	5	5
Comparison.Contrast	1,612	82	127
Contingency.Cause	3,376	120	197
Contingency.Pragmatic cause	56	2	5
Expansion.Alternative	153	2	15
Expansion.Conjunction	2,890	115	116
Expansion.Instantiation	1,132	47	69
Expansion.List	337	5	25
Expansion.Restatement	2,486	101	190
Temporal.Asynchronous	543	28	12
Temporal.Synchrony	153	8	5
Total	12,930	515	766

Table 2: The distribution of relation labels in the Most-used Split dataset.

is to use sections 2–21 for the training set, section 22 for the development set and section 23 for the test set (Lin et al., 2009; Rutherford et al., 2017). However, Shi and Demberg (2017) argued that the standard test set was too small for a reliable evaluation especially when second-level classification was employed.

Table 1 shows the distribution of relation labels in the Cross Validation dataset. Note that although we tried to replicate the procedures described by Shi and Demberg (2017) as closely as possible, there remained slight differences in the discourse relation distribution. For comparison, we also trained the model on the Most-used Split dataset. Table 2 shows the relation label distribution in this dataset. We can confirm that the test set distribution diverged from the development set distribution.

4.1.2 Model Configurations

As word embeddings, we used an off-the-shelf word2vec model specified by the CoNLL 2016 Shared Task organizers. Word embeddings were fixed during training except for some words such as special tags <Arg1> and </Skip>.

As we described in Section 3.2, we used `ConceptNet` and coreference resolution (`Coref`) as external knowledge. Note that if none of the external knowledge is used, MAGE-GRU is reduced to vanilla GRU. For ConceptNet, we removed some triplets that contained stop words. We selected all relation types that appeared in the PDTB except `RelatedTo`. The number of relation types is 35. We added reverse relation types for 30 asymmetric relation types (e.g. `AtLocation` and `Causes`). As a result, the number of ConceptNet relation types used in the experiments was increased to 65.

Although ConceptNet was a relatively high-quality and high-coverage knowledge base, it nevertheless (1) contained questionable triplets (e.g., (time, `Antonym`, year)) and (2) failed to cover some important relations (stock, `AtLocation`, market). We mitigated the first problem by checking weights ConceptNet assigned to triplets. We removed triplets whose weight was smaller than 1.0. The second problem might potentially be addressed graph embedding techniques (Xie et al., 2017), but in the experiments, we used a simpler method. For each word in the input, we prepared the top-10 nearest neighbors in terms of cosine similarity of word2vec vectors with the threshold value of 0.6. We searched ConceptNet for all combinations of the original words and neighbors. As a result, the average number of edges given to an argument pair increased from 4.0 to 17.6.

As for a coreference resolution system, we used Stanford CoreNLP[3] (ver.3.7.0). CoreNLP had three different coreference systems. We chose a neural model since it performed the best among the three.

We tested two input formats: `Args` and `Paragraph`. In `Args`, an input sequence started with <Arg1> and ended with </Arg2>. In the longer `Paragraph` format, it started at the beginning of the paragraph that contained `Arg1` and `Arg2`.

Table 4 summarizes the model configurations. We found that mini-batch did not work for our model. Given that training set accuracy was as low as development set accuracy, we conjecture that training

[3] `https://stanfordnlp.github.io/CoreNLP/`

Relation	Count
Synonym	37,689
FormOf	32,547
IsA	27,251
DerivedFrom	12,668
Antonym	10,632
SimilarTo	6,560
DistinctFrom	6,484
HasContext	6,092
AtLocation	5,584
UsedFor	3,287

Table 3: Top 10 ConceptNet relation types, sorted by the frequency counts in the Penn Discourse Treebank.

Description	Values
input	`Args` or `Paragraph`
word embeddings	100,000 words (300 dims)
optimizer	AdaGrad (Duchi et al., 2011)
pooling	summation
hidden layer	1 layer (600 dims)
external knowledge	coreference and/or ConceptNet (10 dims per relation type)
mini-batch size	1
early stopping	yes

Table 4: Configuration of our model.

	Dev	Test
Args	0.3946	0.3820 (±0.013)
+Coref	0.3939	0.3817 (±0.009)
+ConceptNet	0.4000	0.3926 (±0.012)
+Coref+ConceptNet	0.4037	0.3926 (±0.016)
Paragraph	0.3951	0.3814 (±0.010)
+Coref	0.3919	0.3839 (±0.016)
+ConceptNet	0.4060	**0.3980 (±0.009)**
+Coref+ConceptNet	**0.4083**	0.3938 (±0.011)
feedforward (Rutherford et al., 2017)†	-	0.3660 (±0.011)
LSTM (Shi and Demberg, 2017)	-	0.3444 (±0.014)
+Modality	-	0.3767 (±0.018)

Table 5: Accuracy in the Cross Validation dataset. Test result indicates the mean accuracy across folds and the standard deviation. † denotes our reimplementation.

signals in a mini-batch might have canceled out each other.

4.1.3 Models for Comparison

For comparison, we collected model scores from the literature. As for the Cross Validation dataset, we compared the proposed model with that of Shi and Demberg (2017) and a feedforward network (Rutherford et al., 2017). Shi and Demberg (2017) used an LSTM-based model, optionally with surface features derived from Brown Clusters, Modality, etc. The best score was achieved by LSTM+Modality. We reimplemented the feedforward model of Rutherford et al. (2017) because their evaluation was not based on the Cross Validation dataset.

For the Most-used Split dataset, the models for comparison were a feedforward network (Rutherford et al., 2017), a maximum entropy classifier (Lin et al., 2009) and a CNN-based model (Qin et al., 2017).

4.2 Results

Table 5 shows the results for the Cross Validation dataset. In all configurations, our model outperformed Shi and Demberg (2017)'s models. The best score was achieved by `Paragraph+ConseptNet`. The performance gain of `Paragraph+ConceptNet` over `Paragraph` was statistically significant ($p = 2.63 \times 10^{-7}, p < 0.01$) while `Paragraph+Coref` was no different from `Paragraph` ($p = 0.463$). The `Paragraph` models consistently outperformed the corresponding `Args` models when the input was augmented with external knowledge. However, the impact of contextual information was not statistically significant ($p = 0.081$ for `+Coref+ConceptNet`).

A breakdown of the performance by discourse relation is shown in `Paragraph+ConceptNet` are

Sense	F_1
Comparison.Concession	0.0000 (±0.000)
Comparison.Contrast	0.2287 (±0.029)
Contingency.Cause	0.4813 (±0.015)
Contingency.Pragmatic cause	0.0000 (±0.000)
Expansion.Alternative	0.0087 (±0.026)
Expansion.Conjunction	0.4466 (±0.023)
Expansion.Instantiation	0.4498 (±0.037)
Expansion.List	0.2203 (±0.091)
Expansion.Restatement	0.3341 (±0.027)
Temporal.Asynchronous	0.2140 (±0.077)
Temporal.Synchrony	0.0000 (±0.000)
Total	0.3980 (±0.009)

Table 6: F_1 score in `Paragraph+ConceptNet`.

	Dev	Test
Args	0.4214	0.3668
+Coref	0.4214	0.3655
+ConceptNet	0.4252	0.3799
+Coref+ConceptNet	0.4233	0.3877
Paragraph	**0.4408**	0.3708
+Coref	0.4330	0.3642
+ConceptNet	0.4350	0.3603
+Coref+ConceptNet	0.4350	0.3655
Rutherford et al. (2017)	-	0.3956
Lin et al. (2009)	-	0.4020
Qin et al. (2017)	-	**0.4465**

Table 7: Accuracy in the Most-used Split dataset.

shown in Table 6. Comparing Table 6 with Table 1, we can see that the model ended up ignoring very-low-frequency relations such as Comparison.Concession and Expansion.Alternative, which appeared less than 300 times in training set. This problem could possibly be mitigated by leveraging an unlabeled corpus to increase the size of training instances (Jiang et al., 2016).

The results for the Most-used Split dataset are shown in Table 7. In this dataset, the proposed method was outperformed by models in the literature. Although the F-measure of `Args+Coref+ConceptNet` was about 2 points higher than that `Args`, the performance varied too inconsistently to draw meaningful conclusions. For this reason, we support Shi and Demberg (2017)'s argument for the need of cross validation for implicit discourse relation classification.

4.3 Discussion

As we have seen in Table 5, ConceptNet brought performance gain. However, it appears to leave much room for improvement. Consider the following examples:

`Ex1`:
Another, "Jeux Sans Frontieres ," where villagers from assorted European countries make fools of themselves performing pointless tasks, *is a hit in France*. **A U.S.-made imitation under the title "Almost Anything Goes" flopped fast**.
(Comparison.Contrast)

`Ex2`:
HOMESTEAD FINANCIAL CORP., Millbrae, financial services concern, annual revenue of $562 million, OTC, said *three of its 17 Bay-area branches were closed yesterday*. **The company expects all branches to reopen today**.
(Comparison.Contrast)

In `Ex1`, a baseline model wrongly chose Expansion.Conjunction but our model seems to have suppressed the discourse relation presumably because it found the antonym pair "hit" (`Arg1`) and "flopped" (`Arg2`) in ConceptNet. However, our model misclassified `Ex2` as Expansion.Conjunction even though "yesterday" and "today" were correctly identified as antonyms. This indicates that our model might not have given due weight to ConceptNet, possibly because of some noise in the knowledge base.

In our experiments, coreference resolution did not help implicit discourse relation classification. What we relied on was standard pronominal and nominal coreference resolution, but the following example suggests the need for resolving event coreference (Lu et al., 2016):

	Test
Args+Coref+ConceptNet	0.7723 ($\pm$0.008)
Paragraph+Coref+ConceptNet	0.7921 ($\pm$0.006)

Table 8: Accuracy of explicit discourse relation classification. Result indicates the mean accuracy across folds and the standard deviation.

Is an American Secretary of State seriously suggesting that the Khmer Rouge should help govern Cambodia? *Apparently so*. **There are no easy choices in Cambodia, but we can't imagine that it benefits the U.S. to become the catalyst for an all-too-familiar process that could end in another round of horror in Cambodia**.
(Comparison.Contrast)

In this example, `Arg1` is surprisingly uninformative. In order to classify the discourse relation between `Arg1` and `Arg2`, the system would need to identify what "so" in `Arg1` refers to.

To explore the effect of contextual information in detail, we compared the two input formats, `Args` and `Paragraph`, in the task of *explicit* discourse relation classification. The experimental settings are basically the same as in Section 4.1.2, but the Cross Validation dataset now contained explicit relations. The result is shown in Table 8. Contextual information did help in explicit discourse relation classification, with statistical significance at $p < 0.01$.

It is hard to see exactly why we obtained different results for implicit and explicit relations, but a hint is given by the PDTB annotation itself. The PDTB limits arguments to the minimal text needed to interpret a given relation, and provides *supplementary information* to each of `Arg1` and `Arg2` (named `Sup1` and `Sup2`) (The PDTB Research Group, 2007). Consider the following examples:

That pattern hasn't always held, *but recent strong growth in dividends makes some market watchers anxious*. Payouts on the S&P 500 stocks rose 10% in 1988, according to Standard & Poor's Corp., and Wall Street estimates for 1989 growth are generally between 9% and 14%. Many people believe the growth in dividends will slow next year, although a minority see double-digit gains continuing. **Meanwhile, many market watchers say recent dividend trends raise another warning flag: While dividends have risen smartly, their expansion hasn't kept pace with even stronger advances in stock prices**
(Expansion.Conjunction)

The underlined text span is annotated with `Sup1`. This tag is assigned to a text span supplementary to `Arg1` if it appears relevant but not necessary for interpretation. According to Prasad et al. (2014), only 126 implicit relations were annotated with supplementary information while 1,571 explicit relations were. They amounted to about 0.8% and 8% of the whole implicit and explicit relations, respectively. This great gap indicates that explicit relation classification may benefit more from the text chunks outside of the arguments than implicit relation classification. In fact, a baseline model wrongly chose Comparison.Contrast in this example, but our model chose a correct discourse relation. It should be noted that according to Prasad et al. (2014), consistency control over supplementary information annotation was rather weak. They warned that the gap could be an accidental feature of the PDTB annotation. However, our results lend support to the hypothesis that the gap reflects an intrinsic feature of the discourse relations, or at least that of the PDTB's task specifications.

5 Conclusion

In this paper, we adopted MAGE-GRU to efficiently incorporate external knowledge into the task of implicit discourse relation classification. The experiments show that external knowledge improved accuracy in this task. In addition to a pair of arguments, the text chunk that preceded the pair in the paragraph was given to the model with the hope that it could help classifying its relation. The contextual information

yielded a significant improvement not for implicit discourse relations but for explicit discourse relations. Additionally, we reconfirmed the need for cross validation in this task, as argued by Shi and Demberg (2017).

In the future, we would like to work on extending the neural network architecture. The high composability of MAGE-GRU means that it can easily be combined with other neural network models that are built on top of RNNs. A bidirectional extension to MAGE-GRU may be worth trying. Another future direction is to look for different sources of external knowledge. The candidates include other knowledge bases (e.g. Freebase (Bollacker et al., 2008)) and results of high-level NLP analyses (e.g. event coreference).

References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA.

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 (Volume 1: Long Papers)*, pages 1726–1735.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151.

Bhuwan Dhingra, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2017. Linguistic knowledge as memory for recurrent neural networks. *arXiv*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association of Computational Linguistics*, 3:329–344.

Kailang Jiang, Giuseppe Carenini, and Raymond Ng. 2016. Training data enrichment for infrequent discourse relations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2603–2614.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv*.

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.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1233.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156.

Jing Lu, Deepak Venugopal, Vibhav Gogate, and Vincent Ng. 2016. Joint inference for event coreference resolution. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3264–3275, December.

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

Todor Mihaylov and Anette Frank. 2016. Discourse relation sense classification using cross-argument semantic similarity based on word embeddings. In *Proceedings of the CoNLL-16 shared task*, pages 100–107.

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 26*, pages 3111–3119.

Masataka Ono, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 984–989.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K. Joshi. 2008. Easily identifiable discourse relations. Technical report, University of Pennsylvania Department of Computer and Information Science.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC2008)*, pages 2961–2968.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.

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: Technical Papers*, pages 1914–1924.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric 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, July.

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)*, pages 256–262, July.

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, April.

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: Human Language Technologies*, pages 799–808.

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*, pages 281–291.

Niko Schenk, Christian Chiarcos, Kathrin Donandt, Samuel Rönnqvist, Evgeny Stepanov, and Giuseppe Riccardi. 2016. Do we really need all those rich linguistic features? a neural network-based approach to implicit sense labeling. In *Proceedings of the CoNLL-16 shared task*, pages 41–49.

Wei Shi and Vera Demberg. 2017. Do we need cross validation for discourse relation classification? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 150–156.

Swapna Somasundaran, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 170–179.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3679–3686.

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

The PDTB Research Group. 2007. The Penn Discourse Treebank 2.0 Annotation Manual. Technical report, Institute for Research in Cognitive Science, University of Pennsylvania.

Jianxiang Wang and Man Lan. 2016. Two end-to-end shallow discourse parsers for English and Chinese in CoNLL-2016 shared task. In *Proceedings of the CoNLL-16 shared task*, pages 33–40.

Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. 2017. An interpretable knowledge transfer model for knowledge base completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 950–962, July.

Nianwen Xue, Tou Hwee Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16.

Nianwen Xue, Tou Hwee Ng, Sameer Pradhan, 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.

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*, pages 2230–2235.