A Regularization Approach for Incorporating Event Knowledge and Coreference Relations into Neural Discourse Parsing

Zeyu Dai, Ruihong Huang
Department of Computer Science and Engineering
Texas A&M University
{jzdaizeyu, huangrh}@tamu.edu

Abstract

We argue that external commonsense knowledge and linguistic constraints need to be incorporated into neural network models for mitigating data sparsity issues and further improving the performance of discourse parsing. Realizing that external knowledge and linguistic constraints may not always apply in understanding a particular context, we propose a regularization approach that *tightly* integrates these constraints with contexts for deriving word representations. Meanwhile, it *balances* attentions over contexts and constraints through adding a regularization term into the objective function. Experiments show that our knowledge regularization approach outperforms all previous systems on the benchmark dataset PDTB for discourse parsing.

1 Introduction

Discourse parsing and identifying rhetorical discourse relations between two text spans (i.e., discourse units, either clauses or sentences) is crucial and beneficial for a wide variety of downstream tasks and applications such as machine translation (Webber et al., 2017), text generation (Mann, 1984; Bosselut et al., 2018) and text summarization (Gerani et al., 2014).

In the PDTB-style discourse parsing (Prasad et al., 2008), we commonly distinguish implicit discourse relations from explicit relations, depending on whether a discourse connective (e.g., "because", "however") appears between two discourse units. In general, recognizing implicit discourse relations is more challenging due to the lack of connective, which has recently drawn significant attention from the NLP researchers.

Recent research for implicit discourse relation classification has mostly focused on applying powerful neural network models (Qin et al., 2016a,b; Liu and Li, 2016; Lei et al., 2017; Bai and Zhao, 2018) for modeling compositional meanings and word-level interactions of two discourse units. More recent research has also exploited utilizing broader contexts (Dai and Huang, 2018) as well as leveraging external training data (Xu et al., 2018). Although progress has been made, the performance of implicit discourse relation identification remains low (macro F1 $< 50\%$).

We believe that the low performance is mainly due to the data sparsity issue (Braud and Denis, 2015), which hinders data-thirsty neural network models from making further improvements. Considering the following example from PDTB with two discourse units (DUs):

***DU1*:** The editorial of the WHO notes that tobacco consumption and **lung-cancer** mortality rates are rising in developing countries.

***DU2*:** "No **smoking** should be established as the norm of social behavior" around the world, the editorial says, through the enactment of laws that limit advertising and promote anti-smoking education.

***Discourse Relation*:** Implicit Contingency.**Cause**

Humans can easily recognize this discourse relation as "Cause" because we know that "smoking" is the key causal factor for "lung-cancer", but it is extremely difficult for neural network models trained with limited amount of data to detect it considering the keyword "lung-cancer" only appears few times in the whole PDTB data.

We further argue that external knowledge and linguistic constraints need to be considered for improving implicit discourse relation classification since human annotators also rely on these commonsense knowledge (e.g., smoking causes the lung-cancer) to label the discourse relations. *First*, we consider external event knowledge, because discourse relations (e.g., cause and temporal relations) are often defined as the relation between two

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2976–2987,
Hong Kong, China, November 3–7, 2019.

events (situations in general) as described in two discourse units. As shown in the above example, the "Cause" discourse relation between the two DUs depends on the relation between two events *"smoking"* and *"lung-cancer"* with one event in each DU. *Second*, we consider entity coreference relations as a useful form of linguistic constraints in inferring discourse relations. This is motivated by prior work (Rutherford and Xue, 2014; Ji and Eisenstein, 2015) showing that coreference based features can improve entity mention representations within a DU, which facilitates recognizing coherence and discourse relations between DUs.

In this paper, we investigate how to incorporate external event knowledge and entity coreference relation based linguistic constraints into neural network models for discourse parsing. One key difficulty we want to address is that external knowledge derived event relations or hard linguistic constraints may not always apply for interpreting a particular context, and may hurt performance if used blindly (Kishimoto et al., 2018). Therefore, we propose to *tightly* integrate these constraints into the discourse relation inference process by manipulating hidden word representations to reflect relations between words, and meanwhile *balance* attentions to contexts and constraints through adding a knowledge regularization term in the final objective function.

Specifically, we choose the paragraph-level model we proposed (Dai and Huang, 2018) as the base model, which exploits wider paragraph-level contexts and has been shown effective for PDTB-style discourse parsing. The model mainly consists of a two-level hierarchical BiLSTMs (Schuster and Paliwal, 1997) for modeling both word-level and DU-level inter-dependencies (with a brief description in section 3.1). To implement the knowledge guided regularization for discourse parsing, we *first* insert a new knowledge layer between the word-level BiLSTM and DU-level BiLSTM layer. This knowledge layer modifies hidden representations of words that participate in an event or coreference relation, by applying a relation type specific feedforward neural network. *Then*, we compose a knowledge regularizer based on word representation outputs of the knowledge layer, by adapting a classic knowledge embedding method TransE (Bordes et al., 2013). The regularization term is added to the overall objective function and minimized during model training.

The experiments on PDTB v2.0 demonstrate that our proposed knowledge regularization approach can effectively utilize several types of externally obtained event knowledge and entity coreference relations[1], and improves the performance of both implicit and explicit discourse relation recognition compared to all previous work.

2 Related Work

2.1 Discourse Parsing on PDTB

With the release of Penn Discourse Treebank (PDTB) (Prasad et al., 2008), the task of discourse parsing, especially implicit discourse relation recognition, has received a lot of attention from the NLP community and researchers (Pitler and Nenkova, 2009; Lin et al., 2014; Xue et al., 2015; Rutherford and Xue, 2016). A large number of previous work attempted to model the semantic meanings of two discourse units using latest and advanced neural network models (Chen et al., 2016; Ji et al., 2016; Rutherford et al., 2017; Qin et al., 2017; Guo et al., 2018; Bai and Zhao, 2018). Paragraph-wide contexts were considered for building better discourse unit representations in Dai and Huang (2018). Another research direction for improving implicit discourse relation classification is to expand the training data by leveraging explicit relations (Liu et al., 2016; Lan et al., 2017) or discourse connective informed unlabeled data (Rutherford and Xue, 2015; Xu et al., 2018).

2.2 Incorporate Knowledge into Discourse Parsing

Only a few previous work (Park and Cardie, 2012; Biran and McKeown, 2013; Lei et al., 2018) has exploited external knowledge, including WordNet features (e.g., Antonyms and Hypernyms) and Verb Class (Levin, 1993), in discourse parsing by deriving discrete indicator features and then feed them into feature-based classifiers. Incorporating knowledge as additional features into neural network models often generalize poorly due to the sparsity of features, as also shown in our experiments. Recently, Kishimoto et al. (2018) incorporated the whole of ConceptNet into a MAGE-GRU (Dhingra et al., 2017) based neural networks, but their experiments show that it did not work well for improving implicit discourse relation identification compared with their own base-

[1]Entity coreference relations were generated using an existing coreference resolver from Standford CoreNLP toolkit.

line. We interpret this negative result as the consequence of using irrelevant (noisy) knowledge types blindly without proper regularization.

There are also recent work (Yang and Mitchell, 2017; Xu et al., 2017; Zhou et al., 2018) that incorporate external knowledge into neural network models for improving several other NLP tasks, including information extraction and conversation generation, which mostly followed the *two-step* approach that first obtained representations of knowledge (with triplet format) from knowledge base using knowledge graph embedding methods such as TransE (Bordes et al., 2013), and then utilized attention mechanism (or added gates in a RNN cell (Ma et al., 2018)) to integrate knowledge representations with hidden word vectors. This approach has two main drawbacks: (1) Knowledge representations learned from the first step are fixed without considering the influences of contexts, which may be suboptimal when used for understanding a particular context. (2) With no filtering or regularization, it is difficult for attention mechanisms to explicitly select and attend to the relevant knowledge. In contrast, our proposed regularization approach can be regarded as an end-to-end *joint-learning* framework for discourse parsing and knowledge representation learning, which not only considers both knowledge and contexts in knowledge-aware word representation learning, but also naturally balances attentions on both contexts and knowledge through regularization.

3 Model

Figure 1 illustrates the overall architecture of our model, which implements our knowledge regularization approach (the right part) on top of an existing model as the base model (the left part). There are only two modifications we made to the base model: (1) we insert a novel knowledge layer between the two BiLSTM layers of the base mode; (2) we add a regularizer into the overall objective function. We will first briefly describe the base model, a replication[2] of our recently proposed paragraph-level discourse parsing model (Dai and Huang, 2018). We will then explain the knowledge layer and knowledge regularizer we added.

[2]In our re-implementation, we made several minor modifications to the original base model by using character-level features as well as supporting both traditional fixed word embeddings (300D GloVe (Pennington et al., 2014)) and latest context-dependent word embeddings (1024D ELMo (Peters et al., 2018)) for word embedding initialization.

3.1 Base Model

The base model processes a paragraph containing a sequence of discourse units each time, and predicts a sequence of discourse relations (both implicit and explicit relations) with one relation between each pair of adjacent discourse units (DU). The base model utilizes a hierarchical BiLSTM to calculate both word-level and DU-level representations, followed by a prediction layer and Conditional Random Field (CRF) layer (Lafferty et al., 2001) for jointly predicting a sequence of discourse relations within a paragraph. The base model consists of the following layers:

Character-level CNN Layer: The character-level features, such as the prefix or suffix of a word, can help alleviate the out-of-vocabulary (OOV) problem and improve the word representation in neural nets (Santos and Zadrozny, 2014). In our base model, we use one layer of CNN[3] with max-pooling to extract character-level representation $\boldsymbol{w}_i^{char}$ for the i-th word of the input paragraph.

Word-level BiLSTM Layer: Given a words sequence $X = (x_1, x_2, ..., x_L)$ as the input paragraph, for each word x_i, we construct the expanded word vector by concatenating its word embedding $\boldsymbol{w}_i^{word}$ with its character-level representation and extra word-level features[4] as:

$$\boldsymbol{w}_i = [\boldsymbol{w}_i^{word}; \boldsymbol{w}_i^{char}; \boldsymbol{w}_i^{features}]$$

The word-level BiLSTM layer will process the sequence of expanded word vectors $(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_L)$ and compute the word x_i's hidden representation at each word index i:

$$\boldsymbol{h}_{x_i} = BiLSTM(\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_L)$$

DU-level BiLSTM Layer: Given the output of word-level BiLSTM $(\boldsymbol{h}_{x_1}, \boldsymbol{h}_{x_2}, ..., \boldsymbol{h}_{x_L})$, we calculate the raw DU representation by applying max-pooling operation (Conneau et al., 2017) over the sequence of word representations for all words within a discourse unit: $\boldsymbol{h}'_{DU_j} = \max_{x_i \in DU_j} \boldsymbol{h}_{x_i}$

Then, the DU-level BiLSTM will process the sequence of raw DU representations and obtain the refined DU representation $\boldsymbol{h}_{DU_j}$ for the j-th discourse unit in a paragraph:

$$\boldsymbol{h}_{DU_j} = BiLSTM(\boldsymbol{h}'_{DU_1}, \boldsymbol{h}'_{DU_2}, ..., \boldsymbol{h}'_{DU_{T+1}})$$

[3]Both character embedding and CNN hidden size is 50.

[4]In this work, we used capitalization (Cap) flag, Part-of-speech (POS) tag and named entity (NER) tag of each word as extra word-level features. The embedding size for Cap/POS/NER is 5/35/20. We used Standford CoreNLP toolkit (Manning et al., 2014) to generate POS and NER tags.

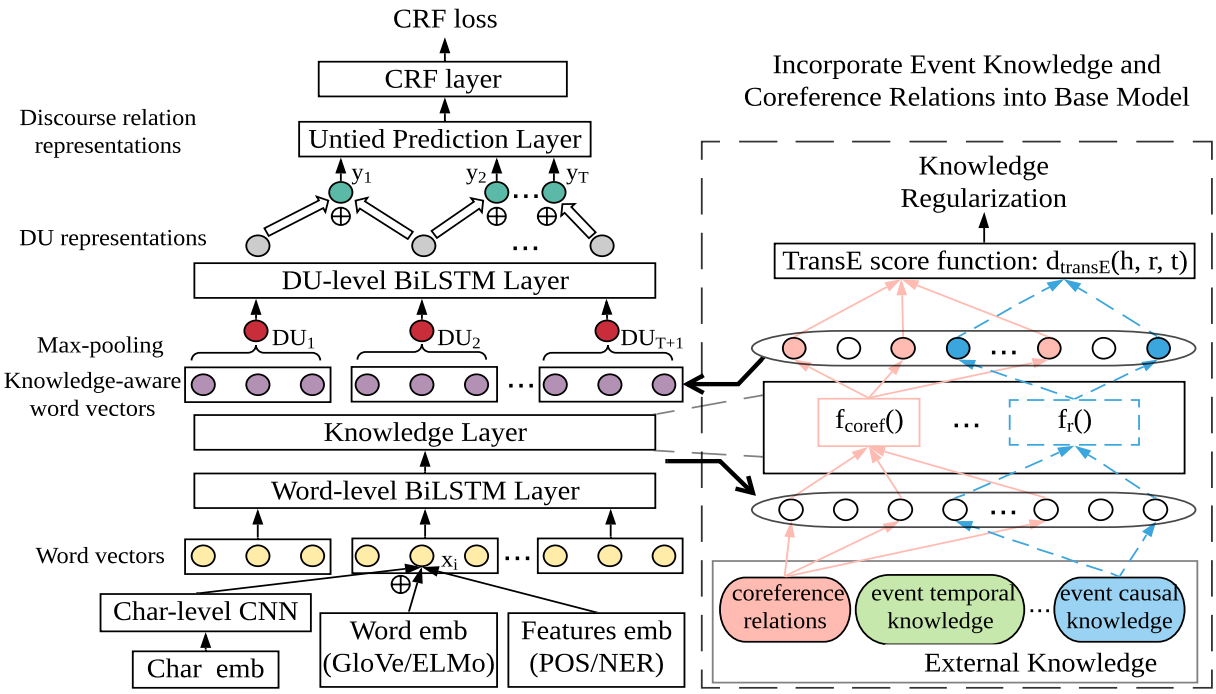

Figure 1: Model Architecture for Paragraph-level Discourse Parsing. The left part is the base model. The right part with colored arrows and neurons show how to incorporate coreference and event knowledge into base model.

Untied (Explicit vs. Implicit) Prediction Layer: Considering the different natures of explicit and implicit discourse relations (Pitler et al., 2009; Lin et al., 2009), the base model trains two independent linear layers with untied parameters for predicting explicit or implicit discourse relations between each two adjacent DUs respectively:

$$\boldsymbol{h}_{y_t} = \begin{cases} W_{exp}[\boldsymbol{h}_{DU_t}; \boldsymbol{h}_{DU_{t+1}}] + b_{exp}, & \text{if } y_t \in exp \\ W_{imp}[\boldsymbol{h}_{DU_t}; \boldsymbol{h}_{DU_{t+1}}] + b_{imp}, & \text{if } y_t \in imp \end{cases}$$

CRF Layer for Discourse Relation Sequence Labeling: A CRF layer (Biran and McKeown, 2015) is added on top of the prediction layer to fine-tune the predicted sequence of discourse relations by capturing continuity and transition patterns (e.g., a temporal relation is likely to follow another temporal relation).

Given the hidden discourse relation representations $\boldsymbol{H}^{(i)} = (\boldsymbol{h}^{(i)}_{y_1}, \boldsymbol{h}^{(i)}_{y_2}, ..., \boldsymbol{h}^{(i)}_{y_T})$ and the target discourse relation label sequence $\boldsymbol{y}^{(i)} = (y^{(i)}_1, y^{(i)}_2, ..., y^{(i)}_T)$ for the i-th training instance, we minimize the following CRF loss function during model training:

$$L_{CRF} = -\sum_i \log p(\boldsymbol{y}^{(i)}|\boldsymbol{H}^{(i)})$$

During testing, the Viterbi algorithm is used to search for the optimal label sequence $\boldsymbol{y}^*$ that maximizes the conditional probability $p(\boldsymbol{y}|\boldsymbol{H})$.

3.2 Knowledge Layer

We simply insert a knowledge layer between the word-level and DU-level BiLSTM layers of the base model, as shown in Figure 1, for incorporating external knowledge and linguistic constraints. Although the knowledge layer can be easily extended to support other types of knowledge, we only consider event knowledge and coreference relations in this paper and leave the exploration of other knowledge types in the future work.

Since there are some notable differences between event relations and entity coreference relations, we model event and coreference constraints in different ways considering their specificities. For example, (1) an event relation can be represented as the triple format $((h, r, t)$ or $(head, relation, tail)$ where $head$ and $tail$ are two event words, $relation$ indicates the relationship between two events $head$ and $tail$.) while a coreference relation can have more than two coreferential entity mentions; (2) event relations are directed while coreference relations are undirected.

Event Knowledge: As the input of our knowledge layer, $\mathbb{E} = (\mathbb{E}_1, \mathbb{E}_2, ..., \mathbb{E}_M)$ denotes the collection of event knowledge triplets generated by matching the paragraph contexts with external event knowledge base (we will give more details in the following section 4.1.), where each triplet has the form of (h, r, t) meaning that there is an event

relation r (either *temporal*, *causal* or *subevent* in this work) between the head event x_h at the position h and tail event x_t at the position t.

For each triplet $\mathbb{E}_m = (h, r, t)$, we use a feedforward neural network[5] $f_r()$ to update the hidden word representations of head and tail events:

$$f_r(\boldsymbol{h}_{x_h}); f_r(\boldsymbol{h}_{x_t}) = \tanh(\boldsymbol{W}_r[\boldsymbol{h}_{x_h}; \boldsymbol{h}_{x_t}] + \boldsymbol{b}_r)$$

where $\boldsymbol{W}_r$ and $\boldsymbol{b}_r$ are relation-specific weights and bias learned for each type of event relation r *only*.

Coreference Relations: Our system assumes that coreference relations in each paragraph are given in the form of coreference clusters, which are generated by running an existing coreference resolver (Clark and Manning, 2016) from the latest version (3.9.2) of Stanford CoreNLP toolkit.

Let $\mathbb{C} = (\mathbb{C}_1, \mathbb{C}_2, ..., \mathbb{C}_K)$ denote coreference clusters in one paragraph, where $\mathbb{C}_k$ contains the word indices with corresponding words referring to the same entity. Similar as above, we use one feedforward neural network $f_{coref}()$ to update the hidden word representation $\boldsymbol{h}_{x_i}$ for words within each coreference cluster. Specifically, the output word vector has the following form:

$$f_{coref}(\boldsymbol{h}_{x_i}) = \begin{cases} \tanh(\boldsymbol{W}_{coref}[\boldsymbol{h}_{x_i}; \boldsymbol{h}_{\mathbb{C}_k}] + \boldsymbol{b}_{coref}), & \text{if } x_i \in \mathbb{C}_k \\ \boldsymbol{h}_{x_i}, & otherwise \end{cases}$$

where $\boldsymbol{W}_{coref}$ and $\boldsymbol{b}_{coref}$ are the weights and bias, $\boldsymbol{h}_{\mathbb{C}_k}$ is a coreference vector calculated by applying max-pooling to all word representations in one cluster: $\boldsymbol{h}_{\mathbb{C}_k} = \max_{x_i \in \mathbb{C}_k} \boldsymbol{h}_{x_i}$. The role of coreference vector is similar to "context vector" utilized in soft attention mechanism (Bahdanau et al., 2015), but we use simple max-pooling instead of computing weights[6] for different word vectors.

3.3 Knowledge Regularization

Inspired by the success of TransE (Bordes et al., 2013) approach in knowledge representation learning, we adapt the key assumption of TransE (i.e., we want $\boldsymbol{h} + \boldsymbol{r} \approx \boldsymbol{t}$ when (h, r, t) holds.) to our framework and hypothesize that the hidden representation of tail t should be close to the hidden representation of head h plus the relation-specific vector $\boldsymbol{h}_r$ in vector space if (h, r, t) holds.

To guide the knowledge-aware word representation learning of the knowledge layer, we propose a knowledge regularizer based on TransE[7] score function $d_{transE}()$ and apply it to the output word vectors of the knowledge layer. The resulting regularization term is also minimized as a part of the objective function during model training. In other words, the knowledge regularization will smoothly penalize constraint depending on whether this constraint can be applied for interpreting relevant relation in a particular context.

Specifically, we use cosine similarity[8] to measure the similarity of two vectors, so the score function for triplet (h, r, t) has the following form:

$$d_{transE}(h, r, t) = 1 - cos(f_r(\boldsymbol{h}_{x_h}) + \boldsymbol{h}_r, f_r(\boldsymbol{h}_{x_t}))$$

where $\boldsymbol{h}_r$ is the relation-specific vector which will be updated as parameters during model training. The event knowledge regularization term is:

$$R_{event} = \sum_{\mathbb{E}_m = (h,r,t) \in \mathbb{E}} d_{transE}(h, r, t)$$

For coreference relations, we create a special triplet $(h, coref, t)$ for each two entity mentions h and t in one coreference cluster, and fix the relation-specific vector $\boldsymbol{h}_{coref}$ to be a zero vector representing the relation of being "identical". The coreference relation regularization term is:

$$R_{coref} = \sum_{\mathbb{C}_k \in \mathbb{C}} \sum_{(h,t) \in \mathbb{C}_k} d_{transE}(h, coref, t)$$

Hence, the overall loss function for our model is:

$$L = L_{CRF} + \lambda_{event} * R_{event} + \lambda_{coref} * R_{coref}$$

4 Experiments

4.1 Dataset and Preprocessing

Dataset: We evaluate our model on PDTB v2.0 (Prasad et al., 2008), which is the largest annotated dataset containing 19K explicit discourse relations and 17K implicit discourse relations. To make our experimental results directly comparable with previous work, we adopted the most-used dataset splitting "PDTB-Ji" (Ji and Eisenstein, 2015) that uses sections 2-20, 0-1, and 21-22

[5] We also tried to use more complicated neural nets including neural tensor networks (Socher et al., 2013) and self-attentions mechanism (Vaswani et al., 2017), but none of them performed better than straightforward feedforward neural network. We even tried to not update (identical function) hidden word vectors, but it performed significantly worse.

[6] We tried to employ weights with soft-attention mechanism, but it did not show improvement in our experiments.

[7] We also tried TransD (Ji et al., 2015) and TransR (Lin et al., 2015) for knowledge regularization, but none of them showed clear improvement over TransE in our experiments.

[8] Note that cosine similarity performed better than L1 or L2 distance in our experiments.

Relation Type	Source	#	Discourse
Coreference	CoreNLP	19,819	Exp & Cont
Event Temporal	Yao et al.	18,515	Temp
Event Causal	ConceptNet	947	Cont
Event Subevent	ConceptNet	626	Exp & Temp

Table 1: Overview of relation types. # is the number of matched triplets (clusters for coreference). The last column summarizes relevant discourse relation classes each knowledge type may potentially help identify.

as train, dev and test sets respectively. To recover the paragraph contexts and gold discourse units, we directly ran the source code[9] of Dai and Huang (2018), and obtained 12,037/1222/1050 paragraph instances in train/dev/test sets respectively.

Knowledge Preprocessing: Table 1 gives an overview of the relation types used in our experiments and the number of triplets (clusters) identified in the PDTB dataset. Specifically, for coreference relations, we utilized the Stanford CoreNLP coreference resolver to identify coreference clusters in each paragraph. For event knowledge, we considered three major event relation types including temporal, causal and subevent. We obtained event temporal knowledge from a previous work (Yao and Huang, 2018)[10] and we retrieved the latter two types of event knowledge from ConceptNet[11] (Speer and Havasi, 2012), which is a widely-used commonsense knowledge base.

4.2 Experiment Setting

Evaluation Setting: Annotated discourse relation labels in PDTB v2.0 are organized in a three-level hierarchy. The top-level coarse-grained discourse relation classes include Comparison (Comp), Contingency (Cont), Expansion (Exp) and Temporal (Temp), which are further split into 16 fine-grained classes at the second-level. To compare with previous work, we report the macro-average F1-score and accuracy[12] on the top-level multi-class classification setting. Note that the macro-average F1-score is normally treated as the main evaluation metric in most previous work considering the imbalanced distribution of discourse relations. In addition, we report class-wise F1-scores for the top-level implicit discourse relations. But different from many previous work that report class-wise F1-scores obtained by using the one-versus-all binary classification setting, we instead report class-wise F1-scores using the 4-way multi-class classification setting, following Dai and Huang (2018) which pointed out that compared to the one-versus-all binary classification setting where all binary classifiers may predict a positive label for one instance, the multi-class classification setting is more appropriate in evaluating a practical end-to-end discourse parser without the need of prediction conflict resolution. We additionally evaluate our models at the second-level using the 11-way[13] multi-class classification.

Training Setting: To make it easy for model tuning, we only chose λ_{coref} and λ_{event} from [0.1, 0.5, 1.0] and tuned them based on the best performance on the dev set. All the BiLSTM layers and our knowledge layer used the hidden state size of 512, so the dimension of all hidden vectors ($\boldsymbol{h}_*$, $f_{coref}(\boldsymbol{h}_*)$ and $f_r(\boldsymbol{h}_*)$) is 512. To prevent gradient exploding problem of LSTMs, we clipped the gradient L2-norm with threshold 5.0 and used L2 regularization with coefficient 10^{-8}. We applied dropout with probability 0.5 on the input/output of BiLSTMs to alleviate overfitting. For the optimizer, we used the SGD with momentum 0.9 and batch size of 64, and we set the initial learning rate as 0.015 which will decay by 5% after each epoch.

To diminish the effects of randomness in neural network model training, we ran all our proposed model, its variants as well as our own base model 3 times using different random seeds and reported the average performance over 3 runs. For fair comparison, we implemented all our models with Pytorch and tested them on a Nvidia 1080Ti GPU.

4.3 Models for Comparison

We compare our proposed regularization models with the following base model, our own baselines and recent published discourse parsing systems:

- (Dai and Huang, 2018): the original model for paragraph-level discourse parsing.

[9] Available at https://github.com/ZeyuDai/paragraph_implicit_discourse_relations. Relations (0.5%) between non-adjacent DUs were discarded.

[10] We also tried to use VerbOcean (Chklovski and Pantel, 2004) which performed worse than our choices.

[11] Available at http://conceptnet.io. To extract event causal knowledge, we merged the relations ['Causes', 'CausesDesire', 'Entails'] defined in ConceptNet. To extract subevent knowledge, we merged the relations ['HasSubevent', 'HasFirstSubevent', 'HasLastSubevent']. For simplicity, we removed relations containing multi-word events or non-event words (e.g., function words).

[12] Note that 3% discourse relations in PDTB were annotated with more than one label. Following previous work (Dai and Huang, 2018; Bai and Zhao, 2018), we considered a prediction as correct if it matches one of the gold labels.

[13] We followed Ji and Eisenstein (2015) to exclude 5 minor second-level classes in our experiments because none of these classes appear in the test or dev sets.

Model	Implicit						Explicit	
	Macro	Acc	Comp	Cont	Exp	Temp	Macro	Acc
Previous work with the same evaluation setting								
(Rutherford and Xue, 2015)	40.50	57.10	-	-	-	-	-	-
(Liu et al., 2016)	44.98	57.25	-	-	-	-	-	-
(Liu and Li, 2016)	46.29	57.57	-	-	-	-	-	-
(Lan et al., 2017)	47.80	57.39	-	-	-	-	-	-
(Dai and Huang, 2018)	48.82	57.44	37.72	49.39	67.45	40.70	93.21	93.98
(Bai and Zhao, 2018) (ELMo)	51.06	-	-	-	-	-	-	-
Our models using GloVe word embeddings								
Base Model	48.96	56.42	41.29	47.77	66.16	40.60	93.78	94.64
Base Model + Coreference (C)	49.42	57.15	41.39	49.26	66.89	40.14	93.73	94.62
Base Model + Event Temporal	49.58	57.31	41.52	46.49	67.50	42.83	93.87	94.72
Base Model + Event (E)	50.02	58.22	40.20	48.06	**68.35**	43.45	93.63	94.46
Full Model (Base Model + C&E)	**50.49**	**58.32**	39.61	**49.29**	68.23	**44.83**	**94.32**	**95.07**
Our own baselines with knowledge features using GloVe word embeddings								
Base Model + Word Features	49.22	57.33	**41.72**	48.30	67.01	39.88	94.05	94.88
Base Model + DU Features	49.27	57.55	**41.72**	48.01	67.36	39.98	93.98	94.82
Base Model + two-step approach	49.63	57.52	41.41	45.57	68.08	43.46	93.88	94.71
Our models using ELMo[16] word embeddings								
Base Model	50.83	56.50	**47.00**	46.42	65.58	44.21	94.35	95.12
Full Model (Base Model + C&E)	**52.89**	**59.66**	45.34	**51.80**	**68.50**	**45.93**	**94.84**	**95.39**

Table 2: Top-level Multi-class Classification Results on PDTB. We report macro-average F1-score (Macro), accuracy (Acc) and F1-score on each class (Comparison, Contingency, Expansion and Temporal). Note that Bai and Zhao (2018) used ELMo word embeddings, so this result is only comparable with the last section.

- Base Model: our replicated model of (Dai and Huang, 2018) for paragraph-level discourse parsing.
- Base Model + Word Features: our own baseline that creates discrete features for each word. We create one feature for each type of relations, including three types of event relations and coreference relations, which counts the number of relation triplets that contain a word. We concatenate these word features with the input word vector $\boldsymbol{w}_i$.
- Base Model + DU Features: our own baseline that creates discrete features for each DU DU_j. We create two features for each type of relations: one counts relation triplets that have both nodes within DU_j; and the other counts relation triplets that have one node in DU_j and the other node in an adjacent DU. We concatenate these DU features with the hidden DU representation $\boldsymbol{h}_{DU_j}$. Adding either word features or DU features is to imitate traditional feature-based approaches and incorporate event knowledge and coreference relation constraints as features.
- Base Model + two-step approach: our own baseline that follows the two-step approach for incorporating relational constraints, including both event relations and coreference relations. We re-implement the inference model proposed by Chen et al. (2018)[14], which first employs TransE to obtain relation representations, and then uses attention mechanism to incorporate relations and build knowledge-enhanced DU representations.
- (Rutherford and Xue, 2015): a feature-based classifier that utilizes explicit discourse connectives for creating more implicit relations.
- (Liu et al., 2016): CNN based multi-task joint learning model that leverages both PDTB and RST (Carlson et al., 2003) datasets.
- (Liu and Li, 2016): a hierarchical attention-over-attention neural network model for implicit discourse relation recognition.
- (Lan et al., 2017): multi-task attention-based model for predicting implicit discourse relations that leverages both explicit discourse relations of PDTB and unlabeled data.
- (Bai and Zhao, 2018): a complex deep residual bi-attention based neural network model for implicit discourse relation classification which improves the hidden representations at character, subword, word and DU levels.

4.4 Experiment Results

Table 2 shows the comparisons. The first section lists the results of previous models that were evaluated on PDTB using top-level multi-class classification. Note that many cells are empty, it is be-

[14]We followed Chen et al. (2018) to tune the hyper-parameter λ for each relation type in the range [0.1, 0.5, 1, 5, 10, 20, 50, 100], and we found that the best result was achieved when λ is set to 0.1 for each type of relations.

	Implicit		Explicit	
Model	Macro	Acc	Macro	Acc
Our models using GloVe word embeddings				
Base Model	28.59	43.88	63.40	86.02
+ Word Features	30.83	44.17	65.47	86.60
+ DU Features	31.78	44.65	66.06	86.51
Full Model	**32.13**	**46.03**	**69.24**	**87.25**
Our models using ELMo word embeddings				
Base Model	31.05	45.98	70.01	87.83
Full Model	**33.41**	**48.23**	**70.48**	**88.08**

Table 3: Second-level Multi-class Classification Results on PDTB. We report macro-average F1-score (Macro) and accuracy (Acc).

cause that many previous publications chose to report the class-wise implicit relation prediction performance using the one-versus-all binary classification setting, which are not directly comparable with our class-wise results using the multi-class classification setting following our previous work Dai and Huang (2018). In addition, we also report the explicit relation results.

In the second section, the replicated model (Base Model) achieves an overall similar performance[15] compared to the original model of Dai and Huang (2018). By incorporating coreference relations (+ Coreference) into the base model using our regularization approach, implicit discourse relation prediction performance was improved for two classes, Contingency and Expansion. Adding event temporal knowledge (+ Event Temporal) into the base model significantly boosts the performance of Temporal discourse relation identification. Furthermore, adding the additional two types of event knowledge (+ Event), causal and subevent relations, yields clear performance gains in predicting another two classes of implicit discourse relations: Contingency and Expansion. These performance gains meet our expectations that event relations are correlated with discourse relations and event relational knowledge facilitates predicting corresponding discourse relations, with correspondences listed in Table 1. The full model considering both event knowledge and coreference relations (+ C&E) achieves further improvements on implicit relation prediction, and outperforms the base model by 1.5 and 1.9 points on macro-average F1-score and accuracy respectively. Meanwhile, our full model obtains the best results for explicit relation prediction as well.

Shown in the third section, our own baselines which incorporate relational constraints either as features or via the two-step approach only perform slightly better than the base model, but clearly worse than the full model using the regularization approach. The first two baselines incorporate constraints as additional discrete features and may suffer from feature sparsity issues, while the two-step approach may fail to balance attentions to contexts and knowledge constraints.

The last section presents the performance when using ELMo word embeddings. Our full model outperforms the base model on three out of four (except Comp) implicit discourse relations and improves both macro-average F1-score and accuracy by 2.1 and 3.2 points respectively. Furthermore, our full model outperforms the previous best system (Bai and Zhao, 2018) using ELMo by over 1.8 points of macro-average F1-score.

[16]We downloaded the pretrained ELMo embedding (5.5B version) from AllenAI's website (https://allennlp.org/elmo) and froze its parameters during model training.

[15]The performance changes on individual categories are due to minor modifications we made in replication, such as adding char-level CNN and replacing word2vec with GloVe.

	Implicit		Explicit	
Model	Macro	Acc	Macro	Acc
Our models using GloVe word embeddings				
Full Model	50.49	58.32	94.32	95.07
w/o Reg	44.28	55.43	92.56	93.59
Our models using ELMo word embeddings				
Full Model	52.89	59.66	94.84	95.39
w/o Reg	48.55	56.45	93.77	94.65

Table 4: Impact of the Knowledge Regularization. We report the top-level performance when knowledge regularizer was removed (w/o Reg) from full model.

5 Analysis

5.1 Second-level Multi-class Classification

Table 3 reports the performance of our models for predicting second-level fine-grained discourse relations. Same as top-level, our full model consistently outperforms the base model and its variants using either word-level or DU-level features.

5.2 Impact of the Knowledge Regularization

To study the necessity of the knowledge regularization in our full model, we removed the regularization terms from the objective function by setting λ_{coref} and λ_{event} to be 0, which essentially means that we did not restrict or regulate the hidden knowledge-aware word vectors at all. From Table 4, we can see that the model without knowledge regularizer performs significantly worse than the full model and even worse than the base model, which supports our hypothesis that using external knowledge or linguistic constraints

Model	# of errors
Base Model	217
Full Model	193

Table 5: Number of errors made by both models for predicting implicit discourse relations with constraints.

blindly can hurt the performance. We conclude that the knowledge regularizer plays a key role in achieving the state-of-the-art performance and the knowledge layer must be used together with knowledge regularization in our framework.

5.3 Qualitative Analysis

To better understand the strengths and weaknesses of the regularization approach, we analyzed implicit discourse relation predictions made by our base model and full model on the dev set. In total, there are 507 implicit discourse relations that match with at least one event or coreference constraint across two DUs, while the remaining 717 instances do not involve those constraints. It turns out that both the full model and base model performed comparably on recognizing implicit discourse relations without event or coreference constraints, with 407 vs. 402 discourse relations correctly predicted by the full model and the base model respectively. Therefore, the overall performance gains achieved by the full model are mainly from better resolving implicit discourse relations with constraints, and as shown in Table 5, the full model made 24 less errors than base model for predicting such implicit discourse relations.

We further compared predictions made by the two models for implicit discourse relations with constraints. We found that 96 predictions in total have been changed by the full model, with clearly more corrections (60, i.e., the full model corrected predictions that were made wrongly by the base model.) than false reversions (36, i.e., the correct predictions made by the base model were wrongly reverted by the full model.). Here is one example from the 60 corrections made by the full model:

***DU1*:** Steve and his firm still **worth** a lot of money.

***DU2*:** A package of credit **support** was put together including the assets of Steve and his firm.

***Gold Discourse Relation*:** Implicit Contingency

***Base Model's prediction*:** Implicit Expansion

***Full Model's prediction*:** Implicit Contingency

The event causal relation between "worth" and "support" identified using external event knowledge has enabled the full model to correctly recognize this Contingency discourse relation.

We further examined the 36 wrong reversions of decisions. Around one third of these errors were due to either noise of event relation knowledge or incorrect coreference relations produced by the external CoreNLP coreference resolver we used. The remaining errors came from over-reliance of the full model on constraints in general. Considering the following example from the 36 reversions:

***DU1*:** Another analyst thought that India may have **pulled back** because of the concern over the stock market.

***DU2*:** India may have felt that if there was a severe **drop** in the stock market and it affected sugar, it could buy at lower prices.

***Gold Discourse Relation*:** Implicit Expansion

***Base Model's prediction*:** Implicit Expansion

***Full Model's prediction*:** Implicit Temporal

The full model could have relied on the event temporal relation between "pulled back" and "drop" and made the wrong discourse relation prediction.

6 Conclusion and Future Work

We have presented an effective regularization approach for incorporating external event knowledge and system predicted coreference relations into an existing paragraph-level neural network model for discourse parsing. Our approach tightly integrates knowledge and linguistic constraints with contexts for deriving knowledge-aware word vectors and meanwhile balances attentions over context and constraints through regularization, which robustly improves both implicit and explicit discourse relation classification performance on the benchmark PDTB corpus. In the future, we will identify new types of commonsense knowledge for further improving the performance of discourse parsing. For example, antonyms (e.g., *warm vs. cold*) can directly indicate a contrast relation between two situations, and this type of knowledge has potential to further improve the performance on *Comparison* discourse relations.

Acknowledgments

We gratefully acknowledge support from National Science Foundation via the award IIS-1755943 and support from Institute of Education Sciences via the award R305A180319.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *COLING*, pages 571–583.

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *ACL*, volume 2, pages 69–73.

Or Biran and Kathleen McKeown. 2015. Pdtb discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 96–104.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NeurIPS*, pages 2787–2795.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *NAACL-HLT*, pages 173–184.

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *EMNLP*, pages 2201–2211.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

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 *ACL*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. 2018. Neural natural language inference models enhanced with external knowledge. In *ACL*, pages 2406–2417.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *EMNLP*.

Kevin Clark and Christopher D Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *ACL*, volume 1, pages 643–653.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, pages 681–691.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *NAACL*, volume 1, pages 141–151.

Bhuwan Dhingra, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2017. Linguistic knowledge as memory for recurrent neural networks. *arXiv preprint arXiv:1703.02620*.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *EMNLP*, pages 1602–1613.

Fengyu Guo, Ruifang He, Di Jin, Jianwu Dang, Longbiao Wang, and Xiangang Li. 2018. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *COLING*.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *ACL*, volume 1, pages 687–696.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *TACL*, pages 329–344.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. In *NAACL-HLT*, pages 332–342.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2018. A knowledge-augmented neural network model for implicit discourse relation classification. In *COLING*, pages 584–595.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001, pages 282–289.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *EMNLP*, pages 1310–1319.

Wenqiang Lei, Xuancong Wang, Meichun Liu, Ilija Ilievski, Xiangnan He, and Min-Yen Kan. 2017. Swim: A simple word interaction model for implicit discourse relation recognition. In *IJCAI*, pages 4026–4032.

Wenqiang Lei, Yuanxin Xiang, Yuwei Wang, Qian Zhong, Meichun Liu, and Min-Yen Kan. 2018. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *AAAI*.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *EMNLP*, 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(2):151–184.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *EMNLP*, pages 1224–1233.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *AAAI*, pages 2750–2756.

Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *AAAI*.

William C Mann. 1984. Discourse structures for text generation. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, pages 55–60.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*, volume 1, pages 2227–2237.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *ACL*, pages 683–691.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *ACL-IJCNLP*, pages 13–16.

R. Prasad, N. Dinesh, Lee A., E. Miltsakaki, L. Robaldo, Joshi A., and B. Webber. 2008. The Penn Discourse Treebank 2.0. In *lrec2008*.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016a. Implicit discourse relation recognition with context-aware character-enhanced embeddings. In *COLING*, pages 1914–1924.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016b. A stacking gated neural architecture for implicit discourse relation classification. In *EMNLP*, pages 2263–2270.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P. Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *ACL*, pages 1006–1017.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *EACL*, pages 281–291.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *EACL*, pages 645–654.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *NAACL-HLT*, pages 799–808.

Attapol T Rutherford and Nianwen Xue. 2016. Robust non-explicit neural discourse parser in english and chinese. *ACL*, page 55.

Cicero D Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *ICML*, pages 1818–1826.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *NeurIPS*, pages 926–934.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann. 2017. Proceedings of the third workshop on discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*.

Yang Xu, Yu Hong, Huibin Ruan, Jianmin Yao, Min Zhang, and Guodong Zhou. 2018. Using active learning to expand training data for implicit discourse relation recognition. In *EMNLP*, pages 725–731.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2017. Incorporating loose-structured knowledge into conversation modeling via recall-gate lstm. In *IJCNN*, pages 3506–3513. IEEE.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi PrasadO Christopher Bryant, and Attapol T Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. *CoNLL*, page 1.

Bishan Yang and Tom Mitchell. 2017. Leveraging knowledge bases in lstms for improving machine reading. In *ACL*, volume 1, pages 1436–1446.

Wenlin Yao and Ruihong Huang. 2018. Temporal event knowledge acquisition via identifying narratives. In *ACL*, volume 1, pages 537–547.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.