Incorporating Image Matching Into Knowledge Acquisition for Event-Oriented Relation Recognition

Yu Hong Yang Xu Huibin Ruan Bowei Zhou Jianmin Yao Guodong Zhou*
School of Computer Science and Technology, Soochow University
No.1 Shizi ST, Suzhou, China, 215006
{tianxianer, andreaxu41, hbr416, zoubowei}@gmail.com;
{jyao, gdzhou}@suda.edu.cn

Abstract

Event relation recognition is a challenging language processing task. It is required to determine the relation class of a pair of query events, such as causality, under the condition that there isn't any reliable clue for use. We follow the traditional statistical approach in this paper, speculating the relation class of the target events based on the relation-class distributions on the similar events. There is minimal supervision used during the speculation process. In particular, we incorporate image processing into the acquisition of similar event instances, including the utilization of images for visually representing event scenes, and the use of the neural network based image matching for approximate calculation between events. We test our method on the ACE-R2 corpus and compare it with the fully-supervised neural network models. Experimental results show that we achieve a comparable performance to CNN while slightly better than LSTM.

1 Introduction

Event relation recognition aims to predict the relationship between the query events. Here, an event is defined as a text span (sentence or clause) which describes the occurrence of a genuine event, such as "*The 2nd industrial revolution*". An event-oriented relation recogntion system is required to assign a relation type tag to a pair of query events, such as those defined in Hong et al (2016)'s natural event relation scheme, including causality, temporality, conditionality, etc. Listed below are a pair of related query events, along with the relation need to be predicted:

(1) **Event Instance 1** – *The 2nd industrial revolution* <Cause>
Event Instance 2 – *The killer fog that blanketed London* <Result>
Relation – `Causality` <Need to be predicted>

Recognizing event relations in an automatic way is a challenging task. It is because the query events are selected from different paragraphs in a document or even different documents, so that there is lack of explicit clue and shared context can be used for semantic relation analysis.

Statistics based inference is one of the promising solutions. It performs in a straight-forward manner: i) seeking for the similar events to the query events, ii) surveying the probability distribution of different types of relations over the similar events, and iii) assigning the most widely distributed relation to the query events. In order to fully implement the inference process, we need to address two crucial issues. One is to collect a large set of pairwise event instances whose relations are either previously known or explicitly signaled, such as the ones in (2). The other is to develop effective similarity measurement approaches, so as to retrieve the event instances similar to the query events.

(2) **Query Event 1** – *China's industrial development in the 21st century* <Cause>
Query Event 2 – *Heavy smog alerts issued for Beijing and other cities* <Result>
Relation – `Causality` <Previously Known>

* Corresponding author

Proceedings of the 27th International Conference on Computational Linguistics, pages 177–189
Santa Fe, New Mexico, USA, August 20-26, 2018.

With regard to the first issue, we employ a connective based approach to collect explicitly related events. Current connective-based explicit relation recognition techniques (Pitler and Nenkova, 2009; Wu et al., 2017) have been proven effective in determining the explicit relations, with no less than 93% accuracy, for the sequential text structures connected by a connective. For example, a pair of clauses connected by the connective "*because*" can be determined to have a `causal` relation with a high level of confidence. This allows us to acquire numerous explicitly-related events from texts using a few carefully-selected connectives and simple patterns. Accordingly, we build a large-scale Textual Event Relation Bank (TERB) to support the relation inference.

We focus on the second issue in this paper, measuring the similarity between events. Recently, neural network has been successfully used in event-oriented semantic encoding to some extent (Nguyen and Grishman, 2016; Ghaeini et al., 2016; Feng et al., 2016; Peng et al., 2016; Liu et al., 2017; Chen et al., 2017), yielding substantial performance gains in the related information extraction tasks, such as event extraction (Doddington et al., 2004; Ahn, 2006) and nugget detection (Ellis et al., 2015). Semantic encoding enables the generation of a high-dimensional distributed representation for characterizing an event. So that it prompts semantic learning, computing and understanding at a very deep level. Undoubtedly, this can facilitate the acquisition of the "semantically-similar but pragmatically-different" event instances for the query events, such as those in (2).

Figure 1: Similar visual scenes

However, in our recent research on utilizing semantic similarity calculation, we fail to pass through a bottleneck that, for some query events, there doesn't exist any semantically-similar event instance in a finite dataset (e.g., TERB). It causes that the performance of the state-of-the-art approaches is actually far from what it should be. In order to overcome the problem, we propose to use visual scenes of events for similarity calculation. It is motivated by the fact that some semantically-dissimilar event instances may possess similar visual scenes with the query events. Figure 1 exhibits the scenes of the query events in (1) and that of the event instances in 3), where the visual scenes are similar (compared between the left and the right images), although the textual descriptions of the events are semantically different.

(3) **Event Instance 3** – *Pollution from steel mills blows over residential buildings* <Cause>
Event Instance 4 – *Mask wearing is in fashion* <Result>
Relation – `Causality` <Previously known>

In our experiments, images are taken as the visual scenes. On the basis, we introduce image captions into cross-media semantic matching, with the purpose of mining possible visual scenes by similarity measurement between query events and image captions. In addition, the Convolutional Neural Network (CNN) based image representation is utilized in visual scene matching. Over the ACE-R2, we compare our model with two fully-supervised discourse relation classification models, including Qin et al. (2016)'s CNN and Chen et al. (2016)'s Long-Short Term Memory (LSTM) based Recurrent Neural Network. Experimental results show that our minimally-supervised model slightly outperforms LSTM and obtain comparable performance with CNN.

In the rest of the paper, we overview the related work in section 2. And then we present the methodological framework in section 3, the caption based cross-media semantic matching in section 4 and image matching in section 5. Section 6 will give the TERB establishment method. In section 7, we evaluate the proposed method and analyze the experimental results. We conclude the paper in section 8.

2 Related Work

2.1 Causal and Temporal Event Relation Identification

So far, the study of event relation parsing mainly concentrates on identifying two kinds of relationships, causality and temporality respectively. The early work on causality identification can be traced back to

the use of lexical-syntactic patterns (Girju et al., 2002; Girju, 2003). Soon thereafter, Chang and Choi (2004) revise Girju et al. (2002)'s patterns using lexical pairs (LP) and cue phrases (e.g., *due to*). Besides they generalize the model for binary relation classification using the Bayes theorem. Abe et al. (2008) further use co-occurrence probability as the novel feature. Recently, scholars have made a great effort to model fine-grained causal relations (Inui et al., 2005), exploit the features for classification (Blanco et al., 2008), refine the patterns by syntactic and discourse parsing (Ittoo and Bouma, 2011; Do et al., 2011) and predict relations by semantic network (Radinsky et al., 2012).

Mani et al. (2006) and Lapata and Lascarides (2006) presented the first study on the temporal relation parsing. Both focus on the machine learning approaches. In the past decade, the SemEval (Verhagen et al., 2007; Pustejovsky and Verhagen, 2009) has promoted a great deal of experimental study, including on the grammatical, syntactic, semantic and ordering features (Bethard and Martin, 2007; Hepple et al., 2007; Puşcaşu, 2007; Bethard, 2013; Mirza and Minard, 2015; Caselli et al., 2015; Hashimoto et al., 2015). Meanwhile, temporal relation modeling has been implemented in different ways, such as sequence labeling, Markov logic networks and hybrid systems (Cheng et al., 2007; Min et al., 2007; Yoshikawa et al., 2009; UzZaman and Allen, 2010; Llorens et al., 2010; Velupillai et al., 2015).

2.2 Multi-class Discourse Relation Classification

In April 2006 (Prasad et al., 2007), the Penn Discourse Tree Bank (PDTB) was released as a corpus of discourse-level arguments as well as annotations of explicit and implicit relations. The PDTB relation scheme consists of 4 main relation classes and 16 sub-classes. Since it is released together with the corpus, a great deal of research has focused on the methodologies for multi-class relation classification.

Motivated by Marcu and Echihabi (2002)'s work, in the earlier study, both feature engineering and sophisticated machine learning dominated the field in a large region (Pitler and Nenkova, 2009; Lin et al., 2009; Louis et al., 2010; Park and Cardie, 2012; Rutherford and Xue, 2014).

Recently, it becomes increasingly popular to use neural networks for learning to classify discourse relations (Zhang et al., 2015; Qin et al., 2016; Chen et al., 2016; Qin et al., 2017; Liu and Li, 2016). Typically, Zhang et al (2015) propose a Shallow Convolutional Neural Network (SCNN) model, which is used by Ponti and Korhonen (2017) to identify contingent relation. Qin et al (2016) utilize CNN with a collaborative gated neural network (CGNN) for recognizing implicit relations. Chen et al (2016) develop a Gated Relevance Network (GRN) based on Bidirectional Long-Short Term Memory (Bi-LSTM) framework, combining bilinear model and single layer network for relevance measurement between arguments. Besides, the current work improves the classification performance of implicit relations by expanding the training data with connectives (Rutherford and Xue, 2015; Braud and Denis, 2016; Wu et al., 2017) or combining multiple corpora via multi-task learning (Liu et al., 2016).

3 Relation Inference Engine

In our statistical inference process, TERB is an indispensable knowledge base and needed to be build first. But in this section, we suppose that TERB has been established successfully, and focus on presenting the inference approach. TERB building will be treated as a separate work and presented in section 6.

3.1 Framework and Terminology

The input of the relation inference engine is a pair of query events, while the output is a relation type tag. The inference engine is constituted of three components, including Cross-Media Semantic Matching (CMSM), image matching and inference model (see the framework in Figure 2). In order to facilitate understanding of the following discussion, we first define the terminologies used as below.

Caption is the textual annotation of an image.

Mention is the textual description of an event.

Visual Scene is an image that best describes how an event happens as well as the surroundings. There is no regard to any concrete element or fact (such as who the participants are).

CMSM verifies whether an event mention has the consistent meaning with a scene, or in other word, the mention evokes the perception of the scene. In our method, CMSM plays an important role because

it helps transform the textual representation (i.e., mention) of an event to a visual version (scene), or vice versa. There are two CMSMs of different directions included in the inference process: forward CMSM and backward CMSM. They are methodologically the same, both measuring similarity between captions and mentions. The difference lies in the use of CMSM. The forward CMSM is used to transform mentions into scenes, while the backward CMSM is applied in a scene-to-mention direction.

Image Matching is put forth for calculating scene similarity. It is conducted between the scenes of the query events and that of the events in TERB. The events in TERB which have the similar scenes with the queries will be adopted, along with their explicit relations. They are used as the reference samples.

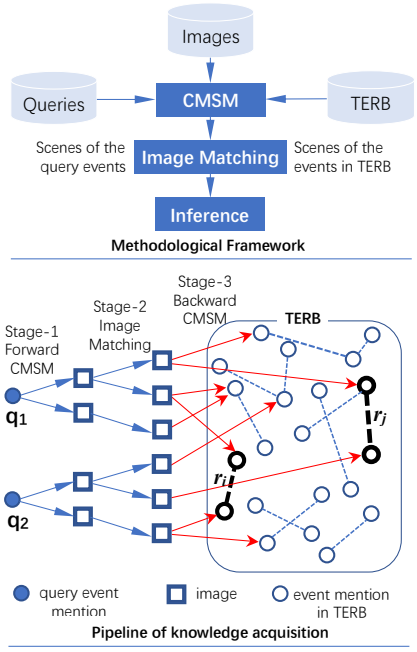

Figure 2: Framework and workflow

3.2 Pipeline for Reference Sample Acquisition

We acquire the reference samples by a three-stage information retrieval system, which consists of three successive one-to-many retrieval stages (see the pipeline in Figure 2):

Stage 1: Mention-to-Scene transformation

The forward CMSM is performed. It is used to retrieve top-n_1 most possible scenes of the query event in a large-scale image database. The images whose captions are most similar to the query event mention will be adopted. The image database (D) we use includes about 5 million images crawled from Wikipedia.

Stage 2: Scene-Scene matching

Each scene obtained in the first stage is used as a query of image search, where image matching is used. There are n_2 most similar images adopted from the image database D, and used as the similar scenes to that of the query event.

Stage 3: Scene-to-Mention transformation

The similar scenes are then used as queries. For each of them, the backward CMSM is performed, so as to acquire n_3 semantically-similar mentions in TERB.

By the knowledge acquisition, for a query event, we can obtain N (N=$n_1 \times n_2 \times n_3$) reference samples in TERB. Thus, given a pair of query events q_1 and q_2, we obtain a collection S of pairs of reference samples (in the size of $N \times N$). Most of the pairs have gone out of use in practice because they are irrelevant and fail to hold a relation. The rest will be reserved as the available reference samples. Their relations are named as reference relations (see r_i and r_j in Figure 2).

3.3 Statistical Inference

Given the set S of reference relations, we use the maximum likelihood estimation to speculate the relation r^* between the query events:

$$r^* = argmax\ p(r) \quad \exists r \in R \qquad p(r) = (\omega_r)^{-1} \times \frac{C(r)}{\sum_{r \in R} C(r)} \qquad \omega_r = \frac{\check{p}(r)}{\lambda_r} \tag{1}$$

where, R denotes all kinds of predefined relation types. We follow Hong et al. (2016)'s relation scheme to discrminate among different relation types. The function $C(r)$ computes the occurrence frequency of the relation r in S. The coefficient ω_r is prodced with a penalty factor λ_r and the prior probability $\check{p}(r)$ of r. The penalty λ_rs of different relation types are inconsistent and need to be fine-tuned on a development set. The prior probability $\check{p}(r)$ of every considered relation type need to be obtained on the training set beforehand. In this paper, we calculate $\check{p}(r)$ using the distribution of r in TERB.

4 Cross-Media Semantic Matching (CMSM)

We implement CMSM between a scene (image) and a mention as the semantic approximation calculation between the caption of the image and the mention. Thus it can be boiled down to a text matching problem.

Activation Function	Loss Function
$tanh_\theta(x) = \frac{e^{\theta^\top x} - e^{\theta^\top x}}{e^{\theta^\top x} + e^{\theta^\top x}}$	$L = -\sum_r y_r \cdot log(s_\theta(x)_r)$
Batch size: 256	**Epoch**: 5
Optimizer: Adadelta **Learning rate**: 10^{-3}	**Dropout ratio**: 0.5 (FC)

Table 1: Hyper-parameter settings for CMSM

Activation Function	Loss Function
$Relu_\theta(x) = ln(1 + e^{\theta^\top x})$	$L = -\sum_r y_r \cdot log(s_\theta(x)_r)$
Batch size: 256	**Epoch**: 74
Momentum: 0.9	**Padding**: 1 pixel
Optimizer: Adadelta **Learning rate**: 10^{-5}	**Dropout ratio**: 0.5 (FC1) **Dropout ratio**: 0.5 (FC2)

Table 2: Hyper-parameter settings for ConvNet

(Note: θ represents all the parameters and $s(*)$ is the ground truth (Vadapalli and Gangashetty, 2016))

For a caption and a mention, we encode each of them as a sentence-level embedding (Sen2vec). We calcualte the cosine similarity of the Sen2vecs and use it as the caption-mention approximation. CNN is utilized for generating the Sen2vecs. As usual, it produces the convolutional features on the concatenated word embeddings of the words in the input short text. There is only 1 hidden layer included in the network. Thus it fails to possess very deep-level perceptions of semantics. Undoubtedly, it can be enhanced by either adding more hidden layers to the network, or instead, using other state-of-the-art network models, such as attention based bidirectional LSTM (Zhou et al., 2016) or gated recurrent unit (Vadapalli and Gangashetty, 2016). In this paper, we choose to use a relatively simple model because we are more willing to verify the validity of the methodological framework.

The configuration of the employed CNN is presented as below: in the **input layer**, a short text (caption or mention) is represented as a fixed-size sequence of real numbers, involving 30 256-dimensional word embeddings. Zero padding is performed when the text length is smaller than 40, otherwise tail clipping. We follow Mikolov et al. (2013) to use skip-gram based word2vec to compute embeddings, and conduct training on the English articles in the latest 2015 Wikipedia dump (Dos Santos and Gatti, 2014). In the **hidden layer**, there are 128 (3×256) filters used for the convolutional computation with a stride of 1. This yields 28 128-dimensional feature vectors. Max pooling is then used to produce a lower dimensional (1×128) vector. As usual, we apply a dense layer to slightly increase the depth. And further, a dropout layer (rate=0.5) is used to produce a 64-dimensional vector. The vector produced in this way is specified as the semantic representation of the text. In the **output layer**, a fully-connected (FC) layer is used, followed by a 18-way softmax classification layer. We train the model on about 50 thousand short texts of 18 domains. Table 1 shows the parameter settings.

It is noteworthy that the domain classification has nothing to do with the task mentioned in this paper. What we really need in the training process is just the well-trained network. The network can be practically used to generate the sentence embeddings (i.e., the ones in FC layer) for captions and mentions.

5 Image Matching

Image matching is used to recognize similar visual scenes and thus enables the events of similar scenes to be acquired. Similarly, we apply the cosine similarity between image embeddings to align images. Simonyan and Zisserman (2014)'s ConvNet is employed to generate the image embeddings.

ConvNet provides an architecture for learning visual features. Each feature indicates a spatial concept of an image patch (only in the input layer) or a multiple-cell receptive field (in hidden layers), preserving the local information about the notions of left, right, top, down and center.

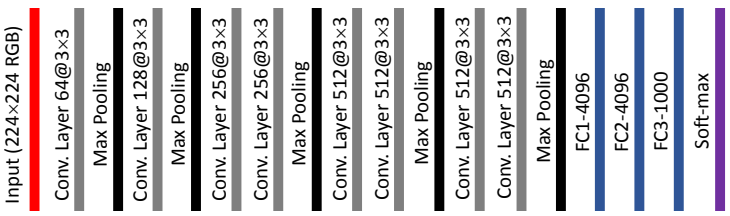

Figure 3: Structure of ImageNet (Plan-A)

In this paper, we follow Simonyan and Zisserman (2014)'s Plan-A, in which a relatively shallower ConvNet is established. The deeper versions in their other plans weren't taken into consideration. It is for the same reason that we mentioned earlier, evaluating the feasibility of our methodological framework when the models we use are weaker than the state of the art.

The ConvNet structure is presented as below and the parameter settings are listed in Table 2. In the **input layer**, a fixed-size 224×224 RGB image is input to the network. The preprocessing we do includes two aspects. One is to normalize images in batches by limiting the size of each to be uniformly 224×224 pixels. The other is to subtract the mean RGB value from each pixel. The mean RGB is computed on the training set. In the **hidden layers**, there are totally 8 convolutional layers deployed, along with 5 max-pooling layers. The number of filters in the convolutional layers is gradually increased (from 64 to 512) with depth. See the layout in Figure 3. In the **output layers**, there are 3 fully-connected (FC.) layers. The first two have 4,096 dimensions each. The third (FC3) is required to perform 1,000-way ILSVRC (Witten et al., 2016) classification during the training procedure and thus contains 1,000 dimensions. A non-linear softmax layer is deployed behind the FCs.

We train the network on ILSVRC-2012. There are about 5 million images used. In our experiments, we pass an image through the well-trained network and employ the FC3 layer as the image embedding. Instead of reproducing the network, we recommend the potential users to use the open source toolkit[1].

6 Textual Event Relation Bank (TERB)

TERB is an event and relation databank we build. There are about 0.74 million pairs of event mentions included in TERB. The mentions are automatically extracted in pairs from Gigaword corpus (Graff and Cieri, 2003). Each pair of events has been exclusively assigned with an exact relation. Below is a standard sample in TERB: <Event Instance 1: "*pollution emission level keeps rising*"> + **causal** relation + <Event Instance 2: "*mask wearing is in fashion*">.

We only take clause-level event mentions into consideration during the process of building TERB. And the eligible clauses for use are limited to the ones that contain a predicate. Syntactic parsing is used for clause segmentation and predicate identification (Björkelund et al., 2010).

There are three patterns used for pairwise mention extraction, including CEE, ECE and EEC, which are distinguishable by means of the position of a connective (C) relative to the neighbor event mentions (E). Below are the examples, where the words in bold font are connectives: $\{CEE-\textbf{\textit{Although}} < Clause1 > , < Clause2 >\}$; $\{ECE- < Clause1 >, \textbf{\textit{but}} < Clause2 >\}$; $\{EEC- < Clause1 >, < Clause2 >, \textbf{\textit{though}}\}$.

The connectives are the words that structurally link sentence constituents. They have been widely used as the reliable clues for explicit semantic relation resolution between text spans. For example, the connective "*because*" frankly signals a causal relation. Nowadays, the utilization of connectives contributes to the high accuracy of explicit relation recognition, reaching a score more than 93% (Pitler and Nenkova, 2009; Wu et al., 2017). Thus, for a pair of event mentions extracted by the patterns, we determine their explicit relation by the one-to-one correspondence between connectives and relation types.

Relational Types	**Num**
Expansion.Conjunctive (Cnj.)	74
Contingency.Conditional (Cnd.)	180
Comparison.Concessive (Cnc.)	1
Temporality.before/after (Baf.)	919
Temporality.during (Dur.)	129
Expansion.Confirming (Cnf.)	19
Contingency.Causal (Cus.)	505
Comparison.Contrastive (Cnt.)	60
Temporality.Equal (Equ.)	107
Coreferential (Cre.)	277

Table 3: Relation scheme

We employ 50 connectives which are elaborately collected from the PDTB corpus (Prasad et al., 2007). The relations they signal have been manually annotated beforehand and double-checked. We filter some ambiguous connectives before use because they generally signal different types of relations. For example, "*since*" may signal a causal relation with the meaning of "*because*", while in some cases, it signals a temporal relation with the meaning of "*from then on*".

7 Experimentation

7.1 Corpus, Settings and Evaluation

Corpus— We use Hong et al. (2016)'s ACE-R2 corpus in our experiments. Table 3 shows the relation scheme, which consists of 5 main types and 10 subtypes. ACE-R2 contains 2,271 pairs of news events.

[1]https://github.com/BVLC/caffe

Each is annotated with a sole relation. The mentions are selected from the Automatic Content Extraction (ACE) corpus (Doddington et al., 2004).

Settings— Both the CNN models and the captioning model that we use in the experiment have been pretrained with external data. What we need to fine-tune is the number of the reference samples (section 3.2), i.e., N, which is produced by the parameters n_1, n_2 and n_3 (N=$n_1 \times n_2 \times n_3$) in the three-stage knowledge acquisition process. A larger value of N will introduce many noises in our inference process. By contrast, a smaller value causes the lack of reference samples and thus the statistical uncertainty.

Evaluation— The methods we concern are evaluated by the metrics of macro-average Precision (Mac-P), Recall (Mac-R) and F-score (Mac-F). For a particular type t, the positive examples are defined as the event pairs that hold the relation t. Thus for t, the precision score is calculated as the ratio of the positive examples in all the output examples of type t. The recall is defined as the ratio of the output positive examples in all the ground-truth ones.

7.2 Compared Methods

Based on the framework mentioned in section 3, we implement an event relation predictor, named as **Holmes**. For the purpose of validating statistically non-random effect, Holmes is compared with the weighted random sampling (**Baseline 1**). By the baseline method, a test sample is most probably determined to hold one of the widely-distributed relations (such as Temporality). The distribution is computed over the development set. Second, Holmes is required to compete with a pure text based approach (**Baseline 2**). In this approach, the statistical inference (section 3.3) is still followed, and similarly to Holmes, the CNN based Sen2vec (section 3) is involved, and devoted to representing the query event mentions and those in TERB. But unlike Holmes, it skips the steps of CMSM and image matching, and instead, directly acquires n_{bas} similar events from TERB by sentence-level embedding similarity.

In addition, we compare Holmes with two discourse classification models, which are based on more general models and trained in a fully supervised fashion. One is Qin et al. (2016)'s **CNN** model. The other is Chen et al. (2016)'s bidirectional LSTM (**Bi-LSTM**) based recurrent model. There are two kinds of performance of the competitors are reported. One is achieved by training the competitors over the standard PDTB corpus (sections 1-20), which is consisted of no more than 6,234 handmade sentence-level argument pairs and relations. The other is achieved by training the competitors over the TERB corpus, a large-scale set of automatically-extracted mentions and explicit relations.

Most of the relation types in PDTB are compatible with those in ACE-R2 and TERB. For example, the `Temporality` has been divided in two subtypes, `Synchronous` and `Asynchronous`, which are compatible with the `Temproality.Equal&During` and `Temporality.Before&After`. However PDTB fails to include the `Coreference` relation. Therefore, it isn't considered in the experiments. Besides, in ACE-R2, the numbers of available instances of the subtypes `Confirming` and `Concessive` are far from large (see Table 3). This makes it difficult to develop Holmes with a high-level confidence. Therefore, the two types aren't taken into consideration too. Thus the total number of available instances in ACE-R2 is 1,974. We use 1,579 as the development set, occupying about 80 percent of all, while the rest (395) for test. We run 5-fold cross validation and report the best performance, while the performance in every validation experiment will be presented in the discussion subsection.

7.3 Experimental Results

Knowledge Acquisition Performance — In the development process, we fine-tune the parameters n_1, n_2 and n_3 for Holmes, and the parameter n_{bas} for the Baseline 2. The NDCG@N metric is used to access the quality of the obtained reference samples (i.e., the retrieved similar events). For the image based retrieval approach in Holmes, N is the product of n_1, n_2 and n_3, while for the text based approach in Baseline 2, N equals to n_{bas}. The high value of NDCG@N implies that most of the N references highly ranked by similarity scores are reliable, holding the ground-truth relation of the query events.

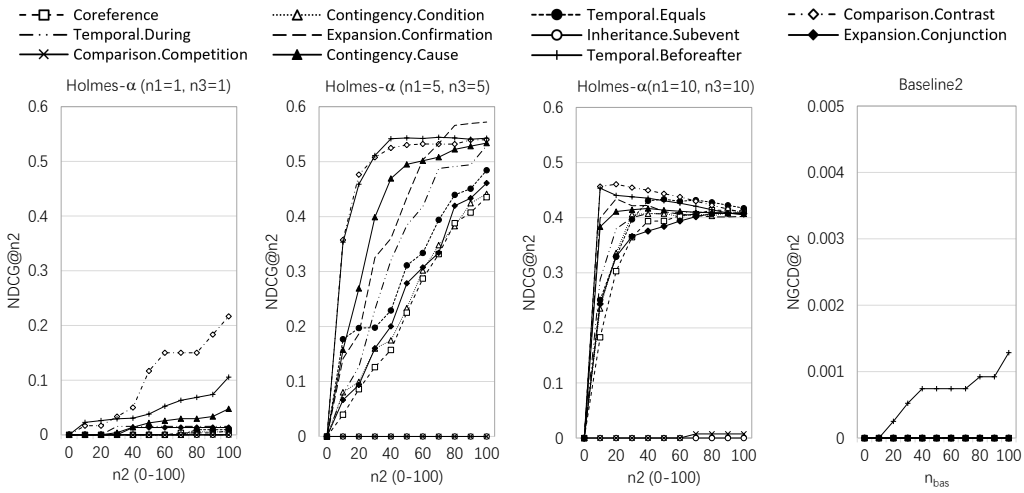

Figure 4: Partial development stages for Holmes-α

Type	Mac-P	Mac-R	Mac-F
Expansion	0.17	0.08	0.11
Contingency	0.82	0.80	0.81
Comparison	0.04	0.07	0.05

Table 4: Best Performance of Holmes for the three main relation types

Subtypes	Mac-P	Mac-R	Mac-F
Contingency.Cus	0.79	0.64	0.71
Contingency.Cnd	0.35	0.53	0.42

Table 5: Results on subtypes of `Contingency`

Subtypes	Mac-P	Mac-R	Mac-F
Temporality.Baf	0.79	0.97	0.87
Temporality.Equ&Dur	0.25	0.04	0.07

Table 6: Results on subtypes of `Temporality`

Figure 4 illustrates partial stages in the development process before the performance remains steady. For example, the subgraph 2 shows the changing tendency of NDCG when n_2 is gradually adjusted, and n_1 and n_3 are kept unchanged. It can be observed from the NDCG curves that the image based approach performs better than the text based (i.e., Baseline 2), obtaining more reliable reference samples. Using such samples, as presented below, we enable Holmes to achieve comparable performance to the sophisticated supervised learning models. It proves that the image based approach performs better than the text based approach in acquiring reference samples. Using such samples, as presented below, we enable the simple statistical inference to outperform some sophisticated supervised learning models.

On the basis, we fine-tune the penalty factor λ for different types of relations. In each-fold cross validation, the penalty factor of a particular relation type will be adopted if it effectively cooperates with the penalty factors of other relation types. The effectiveness is ensured by verifying whether the factors enable the recognizer to achieve the best performance on the development set. We have made the source codes, penalty factors and datasets publicly available[2], so as to enable the reproduction of the whole experiments.

Relation Recognition Performance — We test Holmes by setting the parameters n_1, n_2 and n_3 as 10, 85 and 10. Table 4 shows the performance for the three main relations `Contingency`, `Expansion` and `Comparison`. It is the best performance Holmes achieves in the 5-fold cross validations. Note that the performance for the `Temporal` relation type is evaluated separately and only a 2-way classification is conducted for its subtypes. It is because that every event pair can be regarded to be related temporally, although some of relations fail to be annotated in ACE-R2. The lack of annotations on `Temporality` definitely causes biased assertions on the performance of automated relation prediction.

[2]https://github.com/HuiBinR/VSRB

Method	Training	3-way	Temp	Cont
Baseline1	PDTB	0.11	N/A	N/A
Baseline2	N/A	0.12	N/A	N/A
CNN	PDTB	0.38	0.17	N/A
CNN	**TERB**	0.41	**0.54**	0.60
Bi-LSTM	PDTB	0.39	0.17	N/A
Bi-LSTM	**TERB**	**0.42**	0.48	**0.63**
Holmes	N/A	0.33	0.51	**0.63**

Table 7: Evaluated by General Macro-F

Method	Training	3-way	Temp	Cont
Baseline1	PDTB	0.11	N/A	N/A
Baseline2	N/A	0.12	N/A	N/A
CNN	PDTB	0.36	0.17	N/A
CNN	**TERB**	0.32	**0.48**	0.56
Bi-LSTM	PDTB	0.34	0.17	N/A
Bi-LSTM	**TERB**	**0.40**	0.45	**0.58**
Holmes	N/A	0.31	0.47	0.56

Table 8: Evaluated by Average Macro-F

It is observed that Holmes precisely predicts the `Contingency` relation. Most of the contingently-related instances have been successfully recalled. By contrast, Holmes performs much worse for `Expansion` and `Comparison`. We overview the errors and analyze the reasons as below.

Due to the omission of `Expansion.Confirming`, the performance of Holmes on `Expansion` actually derives from that on `Expansion.Conjunctive`. In general, `Conjunction` appears as a rhetoric method. By `Conjunction`, a series of similar events can be enumerated, such as *earthquake*, *tsunami* and *storm*. Co-occurrence probabilities of such events play important roles for predicting `Conjunction`. However, there are few connectives can be used for pursuing the co-occurred events or they are extraordinarily general. For example, the connective "*and*" is frequently used to signal the co-occurred events, but frankly it nearly connects kinds of linguistic units or even acts as a pause from the perspective of mood. If we use "*and*" to collect sample events which hold conjunctive relations, TERB will be full of pseudo-instances, such as "*sunshine **and** beach*". This will reduce the precision more seriously. By contrast, if we neglect a connective "*and*", there will be lack of available sample events. This inevitably results in the incomplete statistics for the frequently-cooccurred events. We didn't overcome the bottleneck in this paper and use no more than 6 uncommon connectives to collect the sample events. Both the diversity and scale of the events are far from expectation. This makes it difficult to effectively recognize the `Conjunctive` relationship using a statistical approach.

We encounter a similar problem when treating with `Comparison`. A connective like "*than*" is effective to signal a pair of `Contrastive` events. However, due to the limitation of literal expression, in general, such event mentions aren't directly connected by "*than*", but instead the concrete elements in the events are. For example, in a story about *economic competition*, the contrastive events are introduced respectively in different sentences, on the contrary the *profit forecasts* in the events are connected by *than* in a sentence. The neglect of commonly-used connectives (e.g., "*than*") during TERB building causes the lack of sample events. Similarly, the statistic strategy we use fails to perform better under this condition.

Table 5 shows the performance of Holmes on the `Contingent` sub-type relations, i.e., `Cause` and `Condition`. It can be observed that Holmes is adept in predicting causal relations. Table 6 exhibits the performance on the `Temporal` sub-type relations, i.e., `Before&After` and `Equal&During`. It is illustrated that Holmes effectively predicts the synchronous events, though fails to infer the asynchronous.

7.4 Discussion

We carefully evaluate Holmes and the competitors with the general Macro F and average Macro F scores respectively. The former is calculated with the average Macro P and Macro R on the concerned relation types, while the latter is the average value of Macro F on the types. The average Macro F score plays a role of assistance because it is able to reveal the unbalance problem. If a system shows significantly different performance on different types of relations. the average Macro F will be substantially lower than the general. Tables 7 and 8 respectively exhibit the performance of the entrants for the 3 main relation types (3-way), `Temporal` subtypes (Temp) and `Contingent` subtypes (Cont). The tables only exhibit the best performance the competitors achieves through the 5-fold cross validations. The performance in every validation experiment is shown in Figure 5 and 6.

It can be observed that Holmes performs worse than CNN and LSTM for the 3-way main relation classification. Nevertheless, Holmes obtains a smaller gap between general and average Macro F scores. Besides, it achieves a comparable performance to the well-trained CNN and LSTM for the fine-grained sub-type relations. For the case of `Contingency`, Holmes reaches the top together with LSTM. Note

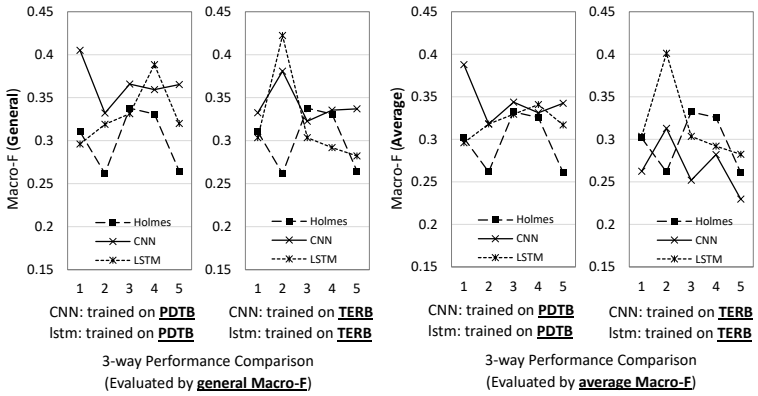

Figure 5: Cross validation for main relation types (`Contingency`, `Comparison` and `Expansion`)

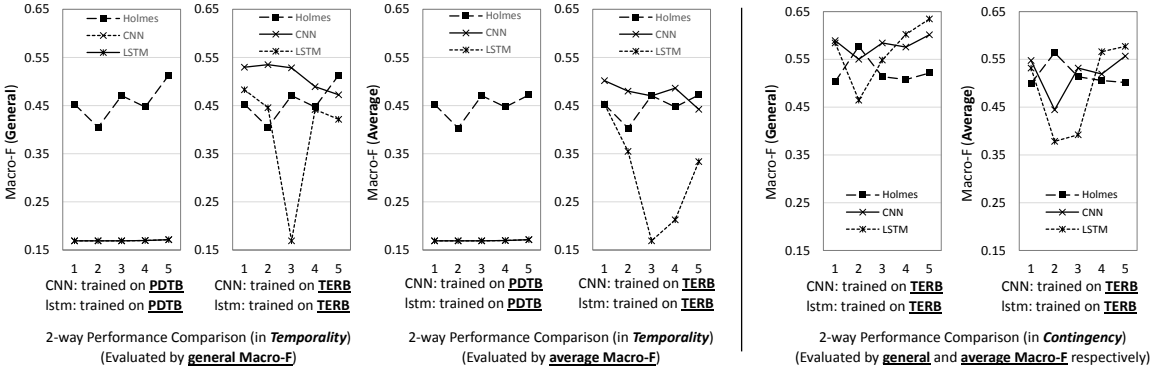

Figure 6: Cross validation for the subtypes in `Temporality` and `Contingency`

that CNN and LSTM fail to be trained on PDTB for classifying the `Contingent` sub-type relations. It is because PDTB contains only one implicit conditionally-related argument pair. The scale of training data and domain adaptation impose great influences on the performance. For example, for the subtypes, CNN and LSTM perform significantly worse than Holmes when trained on PDTB. In addition, Holmes outperforms Baseline 2. It demonstrates that, in TERB, there is lack of semantically-consistent events to the queries. Holmes bypasses the bottleneck, using CMSM and image matching to acquire visually-similar events. This contributes to the reference sample based statistical inference.

When we look through the 5-fold cross validation, we find that Holmes actually performs better than expectation. As shown in Figures 5 and 6, Holmes seldom shows an exceptionally high or an unusually low Marcro F score. And the mean level of Macro F scores of Holmes appears to be no less than that of LSTM for the subtypes of both `Temporality` and `Contingency`.

8 Conclusion

We demonstrate that the use of images also contributes to knowledge acquisition in the field of linguistic computing. By image matching, we open up a new perspective for the identification of similar events, i.e., measuring the similarity of visual scenes. On the basis, we have successfully acquired a variety of events which have comparable scenes with the queries. Using the obtained events and their explicit relations, we enable a simple statistical inference model to achieve competitive performance for event relation recognition. In the future we will introduce active learning into the refinement of the collected event instances, and expand the existing training data to strengthen the supervised classification models.

Acknowledgements

We thank Siyuan Ding and Liang Yao who have made great efforts on this work. This work was supported by the national Natural Science Foundation of China (Nos. 61751206, 61672368, 61672367).

References

Shuya Abe, Kentaro Inui, and Yuji Matsumoto. 2008. Acquiring event relation knowledge by learning cooccurrence patterns and fertilizing cooccurrence samples with verbal nouns. In *IJCNLP*, pages 497–504.

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8. Association for Computational Linguistics.

Steven Bethard and James H Martin. 2007. Cu-tmp: Temporal relation classification using syntactic and semantic features. In *Workshop on Semantic Evaluations*, pages 129–132.

Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics*, volume 2, pages 10–14.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 33–36. Association for Computational Linguistics.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *LREC*.

Chloé Braud and Pascal Denis. 2016. Learning connective-based word representations for implicit discourse relation identification. In *EMNLP*, pages 203–213.

Tommaso Caselli, Antske Fokkens, Roser Morante, and Piek Vossen. 2015. Spinoza vu: An nlp pipeline for cross document timelines. *SemEval-2015*, page 787.

Du-Seong Chang and Key-Sun Choi. 2004. Causal relation extraction using cue phrase and lexical pair probabilities. In *IJCNLP*, pages 61–70. 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 (1)*.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 409–419.

Yuchang Cheng, Masayuki Asahara, and Yuji Matsumoto. 2007. Naist. japan: temporal relation identification using dependency parsed tree. In *Semantic Evaluations*, pages 245–248.

Siyuan Ding, Yu Hong, Shanshan Zhu, Jianmin Yao, and Qiaoming Zhu. 2016. Combining event-level and cross-event semantic information for event-oriented relation classification by scnn. In *China National Conference on Chinese Computational Linguistics*, pages 216–224. Springer.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303. Association for Computational Linguistics.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *LREC*, volume 2, page 1.

Cícero Nogueira Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*, pages 69–78.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. 2015. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. In *Proceedings of TAC KBP 2015 Workshop, National Institute of Standards and Technology*, pages 16–17.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 66–71.

Reza Ghaeini, Xiaoli Fern, Liang Huang, and Prasad Tadepalli. 2016. Event nugget detection with forward-backward recurrent neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 369–373.

Roxana Girju, Dan I Moldovan, et al. 2002. Text mining for causal relations. In *FLAIRS*, pages 360–364.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *ACL*, pages 76–83.

David Graff and C Cieri. 2003. English gigaword corpus. *Linguistic Data Consortium*.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. 2015. Generating event causality hypotheses through semantic relations. In *AAAI*, pages 2396–2403.

Mark Hepple, Andrea Setzer, and Rob Gaizauskas. 2007. Usfd: preliminary exploration of features and classifiers for the tempeval-2007 tasks. In *Workshop on Semantic Evaluations*, pages 438–441.

Yu Hong, Tongtao Zhang, Tim OGorman, Sharone Horowit-Hendler, Heng Ji, and Martha Palmer. 2016. Building a cross-document event-event relation corpus. *LAW X*, page 1.

Takashi Inui, Kentaro Inui, and Yuji Matsumoto. 2005. Acquiring causal knowledge from text using the connective marker tame. *TALIP*, 4(4):435–474.

Ashwin Ittoo and Gosse Bouma. 2011. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *Natural Language Processing and Information Systems*, pages 52–63. Springer.

Maria Lapata and Alex Lascarides. 2006. Learning sentence-internal temporal relations. *J. Artif. Intell. Res.(JAIR)*, 27:85–117.

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: Volume 1-Volume 1*, pages 343–351. Association for Computational Linguistics.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. *arXiv preprint arXiv:1609.06380*.

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

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017. Exploiting argument information to improve event detection via supervised attention mechanisms. 1:1789–1797.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Workshop on Semantic Evaluation*, pages 284–291.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit discourse relations. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 59–62. Association for Computational Linguistics.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *COLING and ACL*, pages 753–760.

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

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

Congmin Min, Munirathnam Srikanth, and Abraham Fowler. 2007. Lcc-te: a hybrid approach to temporal relation identification in news text. In *Workshop on Semantic Evaluations*, pages 219–222.

Paramita Mirza and Anne-Lyse Minard. 2015. Hlt-fbk: a complete temporal processing system for qa tempeval. *SemEval-2015*, page 801.

Thien Huu Nguyen and Ralph Grishman. 2016. Modeling skip-grams for event detection with convolutional neural networks. In *EMNLP*, pages 886–891.

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. Association for Computational Linguistics.

Haoruo Peng, Yangqiu Song, and Dan Roth. 2016. Event detection and co-reference with minimal supervision. In *EMNLP*, pages 392–402.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16. Association for Computational Linguistics.

Edoardo Maria Ponti and Anna Korhonen. 2017. Event-related features in feedforward neural networks contribute to identifying causal relations in discourse. *LSDSem 2017*, page 25.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The penn discourse treebank 2.0 annotation manual.

Georgiana Puşcaşu. 2007. Wvali: Temporal relation identification by syntactico-semantic analysis. In *Workshop on Semantic Evaluations*, pages 484–487.

James Pustejovsky and Marc Verhagen. 2009. Semeval-2010 task 13: evaluating events, time expressions, and temporal relations (tempeval-2). In *Workshop on Semantic Evaluations*, pages 112–116.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 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. *arXiv preprint arXiv:1704.00217*.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *WWW*, pages 909–918.

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

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

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Naushad UzZaman and James F Allen. 2010. Trips and trios system for tempeval-2: Extracting temporal information from text. In *Workshop on Semantic Evaluation*, pages 276–283.

Anandaswarup Vadapalli and Suryakanth V Gangashetty. 2016. An investigation of recurrent neural network architectures using word embeddings for phrase break prediction. In *Interspeech*, pages 2308–2312.

Sumithra Velupillai, Danielle L Mowery, Samir Abdelrahman, Lee Christensen, and Wendy W Chapman. 2015. Blulab: Temporal information extraction for the 2015 clinical tempeval challenge.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 75–80. Association for Computational Linguistics.

Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Changxing Wu, Xiaodong Shi, Yidong Chen, Jinsong Su, and Boli Wang. 2017. Improving implicit discourse relation recognition with discourse-specific word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 269–274.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly identifying temporal relations with markov logic. In *ACL and AFNLP*, pages 405–413.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *EMNLP*, pages 2230–2235.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212.