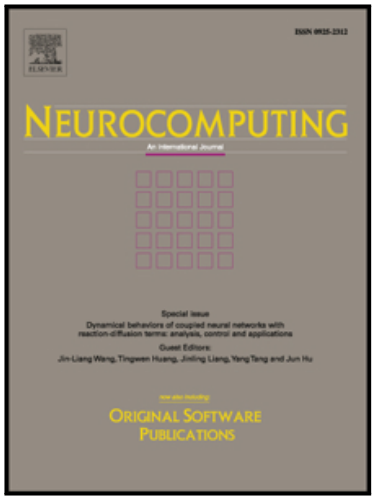
Communicated by Yongdong Zhang

Accepted Manuscript



Learning Better Discourse Representation for Implicit Discourse Relation Recognition via Attention Networks

Biao Zhang, Deyi Xiong, Jinsong Su, Min Zhang

PII: S0925-2312(17)31594-1
DOI: 10.1016/j.neucom.2017.09.074
Reference: NEUCOM 18957

To appear in: *Neurocomputing*

Received date: 28 June 2017
Revised date: 8 September 2017
Accepted date: 24 September 2017

Please cite this article as: Biao Zhang, Deyi Xiong, Jinsong Su, Min Zhang, Learning Better Discourse Representation for Implicit Discourse Relation Recognition via Attention Networks, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.09.074

Learning Better Discourse Representation for Implicit Discourse Relation Recognition via Attention Networks

Biao Zhang[a,c], Deyi Xiong[b], Jinsong Su[a,c,*], Min Zhang[b]

[a] *Xiamen University, Xiamen, China 361005*
[b] *Soochow University, Suzhou, China 215006*
[c] *Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou 350121, Fujian, China*

Abstract

Different words in discourse arguments usually have varying contributions on the recognition of implicit discourse relations. Following this intuition, we propose two attention-based neural networks, namely *inner attention model* and *outer attention model*, to learn better discourse representation by automatically estimating the degrees of relevance of words to discourse relations. The former model only utilizes the information inside discourse arguments, while the latter model builds upon an outside semantic memory to exploit general world knowledge. Both models are capable of assigning more weights to relation-relevant words, and operate in an end-to-end manner. Upon these two models, we further propose a *full attention model* that combines their strengths into a unified framework. Extensive experiments on the PDTB data set show that our model significantly benefits from highlighting relation-relevant words and yields competitive and even better results against several state-of-the-art systems.

Keywords: implicit discourse relation recognition, attention network, memory network, convolutional neural network

1. Introduction

Implicit discourse relation recognition (DRR) is a task of automatically identifying the internal structure and logical relation of a coherent text without discourse connec-

*Corresponding author
Email addresses: zb@stu.xmu.edu.cn (Biao Zhang), dyxiong@suda.edu.cn (Deyi Xiong), jssu@xmu.edu.cn (Jinsong Su), minzhang@suda.edu.cn (Min Zhang)

tives. DRR plays a vital role in a variety of nature language processing applications, such as question answering [1], information extraction [2], machine translation [3] and so on. Although explicit DRR has recently achieved remarkable success [4, 5], implicit DRR remains a serious challenge due to the absence of discourse connectives.

Most existing approaches to implicit DRR heavily rely on various manually designed features to capture rich linguistic information for better performance [6]. In spite of their success, feature engineering in this manner usually requires domain knowledge and the number of hand-crafted features is often huge (especially the combinational features capturing interactions between two arguments). This feature engineering philosophy is rather time-consuming, thus is not easily adaptable.

Concerning this issue, Zhang et al. [7] adopt shallow convolution operations and nonlinear transformations on the top of neural word embeddings to automatically learn input features for implicit DRR. Their model is general and flexible in the sense that it requires no manual engineering. However, one drawback of their model is that it treats all discourse words equally, neglecting the different degrees of importance of words to discourse relations. Let's consider the following implicit discourse relation (comparison):

(1) Our competitors say we overbid them,

(but) who cares.

Obviously, the words "*competitors, overbid, cares*" are more persuasive than the others for relation recognition. Intuitively, discourse relations should be sensitive to words that are tightly related and insensitive to words that occur in any discourse relations. Therefore, a model that can capture and discriminate the discourse-related words would be more decisive when predicting. Previous studies have exploited position-dependent word features, e.g. *First-Last, First3* [6], which manually define different roles of words in DRR. Nevertheless, we prefer developing a model that is able to learn these differences automatically.

Inspired by recent success of attention mechanisms [8, 9, 10] as well as memory networks [9, 11, 12], in this paper, we propose two attention-based neural networks to learn better discourse representations for implicit DRR. The basic idea behind is to treat

different words in discourse arguments unequally and automatically identify different degrees of importance of words to discourse relations so as to make the resulting representation place more emphasis on those discourse-relation-preferred words. According to whether relying on external information, we classify our two models as *inner attention model* and *outer attention model*. The former only requires the discourse argument itself to distinguish the word contributions, which are successively used to weight discourse word embeddings as an alternative of the mean-pooling operation in [7]. In contrast, the latter relies on an additional semantic memory to leverage a more general external world knowledge. Following the spirit of memory network [11], we use the content-based addressing strategy to assign each word a weight so as to retrieve a deep semantic meaning representation for the discourse from the memory. Both kinds of information are complementary to each other, thus we further combine these two models into a unified framework, *full attention model*, to fully access the internal and external information for better discourse representation. All models are end-to-end neural networks, enabling the stochastic gradient descent optimization.

Our major contributions are twofold:

- We propose an *inner*, an *outer* and a *full* attention model to learn better discourse representation for implicit DRR. The exploration of different attention schemes for implicit DRR, to the best of our knowledge, has never been investigated before.
- We conduct a series of experiments for English implicit DRR on the PDTB-style corpus to evaluate the effectiveness of our proposed models.

Experiment results on both one-against-all and four-way classification show that our proposed models yield satisfactory improvements against several strong baselines in terms of F1 score. Extensive analysis on the learned word contributions further discloses some linguistic characteristics of the proposed models.

2. Symbol Definitions

This section provides definitions for basic symbol notations used in this paper, which are shown in Table 1.

Symbol	Definition
Arg_1, Arg_2	The first and second discourse argument annotated in PDTB corpus respectively.
Arg	We use it to denote either Arg_1 or Arg_2.
L, M	The word embedding matrix and semantic memory matrix respectively.
E	We use it to denote either L or M.
X, x_i	The ordered vector list and the vector of the i-th word respectively.
d_1	The dimension of word embedding.
d_2	The dimension of the semantic memory matrix.
d_a	The dimension of the attention space.
c, c_o, c_a	The representation of final, initial and attended discourse argument respectively.
p, p_o, p_a	The representation of final, initial and attended discourse respectively.
v	We use it to denote either discourse or argument representation.
min, max, avg	The min-pooling, max-pooling and mean-pooling operation respectively.
s, α	The semantic matching score and attention weight respectively.
y, y_g	The predicted relation distribution and the ground truth respectively.

Table 1: Basic symbol notations.

3. Background

This section firstly explains the annotations in Penn Discourse Treebank (PDTB) to better understand the DRR problem. We then review how discourse arguments are represented in the shallow convolution neural network (SCNN) model [7] as it forms our basis and the neural baseline.

3.1. Annotations in PDTB

The PDTB [13] annotates discourse relations in a predicate-argument view, where a discourse connective is considered to be a predicate that takes two text spans as its arguments. The argument, to which a discourse connective is syntactically attached, is called *Arg2* (e.g., the second sentence in example (1)). The other argument is called *Arg1* (e.g., the first sentence in example (1)).

Generally, PDTB provides annotations for both explicit and implicit discourse relations. Implicit relations are annotated with connective expressions that best convey

the inferred implicit relations between adjacent sentences within the same paragraph. As shown in example (1), the connective "*but*" is chose to express the inferred COMPARISON relation.

The relation tags in PDTB are arranged in a three-level hierarchy, where the top level consists of four major semantic *classes*: TEMPORAL (TEM), CONTINGENCY (CON), EXPANSION (EXP) and COMPARISON (COM). For each class, a second level of *types* is defined to provide finer semantic distinctions. A third level of *subtypes* is defined for only some types to specify the semantic contribution of each argument. Because the top-level relations are general enough to be annotated with a high inter-annotator agreement and are common to most theories of discourse, in our experiments we use only this level of the annotations.

3.2. Argument Representation in SCNN

In SCNN, each word in vocabulary V is represented as a d_1-dimensional dense, real-valued vector, and all these vectors are stacked into a word embedding matrix $L \in \mathbb{R}^{d_1 \times |V|}$, where $|V|$ is the vocabulary size. Formally, given an argument which is an ordered list of n words, SCNN retrieves vector representation $x_i \in \mathbb{R}^{d_1}$ for each word from L and treats the resulting vector list $\mathrm{X} = (x_1, x_2, \ldots, x_n)$ as its input layer.

SCNN further extracts the major information contained in this vector list for argument representation. This is achieved via several convolution operations *avg*, *min* and *max* as follows:

$$c^{avg} = \text{avg}\,(x_1, x_2, \ldots, x_n) \tag{1}$$

$$c^{min} = \min\,(x_1, x_2, \ldots, x_n) \tag{2}$$

$$c^{max} = \max\,(x_1, x_2, \ldots, x_n) \tag{3}$$

. The discourse argument is represented by the concatenation of these convolutional features, i.e., $c^{Arg} = \left[c^{avg}; c^{max}; c^{min}\right]$, where Arg means the *Arg1* or *Arg2*.

4. Attention-based Neural Networks

4.1. Motivation

Although SCNN subsequently performs several nonlinear transformations on c successively, it treats words in an argument without distinction of their different effects on discourse relations, which may be problematic because discourse relations should be more tightly related to relation-sensitive words as discussed in Section 1. Previous studies that successfully exploit *Verbs*, *First-Last, First3* and *Modality* as features [14] also suggest that words with different parts-of-speech, positions and modalities are different for the discourse relation recognition. We introduce an attention mechanism to model these differences automatically.

4.2. The Method

We aim at improving discourse representations by distinguishing word contributions in discourse arguments for the relation recognition. Formally, instead of assigning each word with an equal weight as in Eq. (1), we weight each word differently

$$\widetilde{c} = \sum_{i=1}^{n} \alpha_i E_i \tag{4}$$

. E_i is the embedding of the i-th word, whose meaning differs in different models. α_i is the attention weight, which should reflect the degree of importance of the word E_i in representing the whole discourse with respect to the final discourse relation recognition. Recall the above-mentioned example (1). If words "*competitors, overbid, cares*" are detected as important words for DRR, there would be more chance that the final recognizer succeeds.

The attention weight α_i is typically generated by normalizing a match score vector over all the words in E,

$$\alpha_i = \frac{exp(s_i)}{\sum_{j=1}^{n} exp(s_j)} \tag{5}$$

with,

$$s_i = g(v, E_i) \tag{6}$$

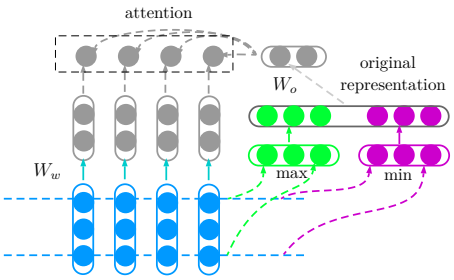


Figure 1: Attention in (S-)InAtt model for one argument. Word embeddings are represented by blue color, while the min and max convolutions are indicated in purple and green color respectively. Additionally, we use gray color to denote the attention space, and nodes in dash box are learned attentional weights.

, where v is the representation of discourse/argument. $g(\cdot)$ is a scoring function that estimates how useful each word E_i is to the corresponding discourse. This scoring function differs in our proposed two models.

The above descriptions can be formulated into the following abstract form:

$$\widetilde{c} = \text{Attention}(v, E) \tag{7}$$

. In this way, the attention network uses the discourse embedding v to extract relation-relevant part $\widetilde{c}$ from words E for relation classification. The overall framework is the same for all our models, but different models may have different inputs v, E and scoring functions $g(\cdot)$. We elaborate our *inner attention model*, *outer attention model* and their combined *full attention model* from this high-level philosophy.

4.2.1. Inner Attention Model

The inner attention model (InAtt) assumes that relation-specific word embeddings are more expressive than the general word embeddings for relation recognition. Therefore, replacing c^{avg} in Eq. (1) with $\widetilde{c}$ in Eq. (4) could benefit the task of implicit DRR. The attention mechanism in InAtt can be summarized as follows:

$$c_a^{Arg} = \text{Attention}(c_o^{Arg}, \text{X}^{Arg}) \tag{8}$$

, where $E = \text{X}^{Arg}$ and $v = c_o^{Arg}$ is the word embedding matrix and original/initial representation for argument Arg respectively. The output c_a^{Arg} is a d_1-dimensional vector for argument Arg alone, with the subscript a denoting the attention.

Figure 1 illustrates the architecture of InAtt. We employ the concatenation of *min* and *max* convolution features as the initial representation for an argument since these features are well exploited and desirable to selectively capture crucial informative aspects inside X^{Arg}:

$$c_o^{Arg} = \left[c^{Arg,max}; c^{Arg,min}\right] \in \mathbb{R}^{2d_1} \tag{9}$$

. This initial representation is further taken to highlight relation-relevant words in Arg via the scoring function $g(\cdot)$ as in [8] (see the gray colors in Figure 1):

$$g(c_o^{Arg}, \mathrm{X}_i^{Arg}) = W_a \tanh(W_o c_o^{Arg} + W_w \mathrm{X}_i^{Arg}) \tag{10}$$

, where $W_o \in \mathbb{R}^{d_a \times 2d_1}, W_w \in \mathbb{R}^{d_a \times d_1}, W_a \in \mathbb{R}^{1 \times d_a}$ are weight matrices and d_a denotes the attention dimensionality. This scoring function firstly projects the word and initial discourse representations into a common feature space, in which it then fully qualifies their semantic matching degrees with non-linear transformation followed by a cosine-style estimation through W_a. If the word X_i^{Arg} is important to relation recognition, the model would tune these weight matrices to automatically align c_o^{Arg} and X_i^{Arg} so as to improve their matching degrees, i.e. a high score $g(c_o^{Arg}, \mathrm{X}_i^{Arg})$. Clearly, the InAtt model only utilizes the information inside discourse arguments.

To represent the whole discourse, InAtt concatenates both the initial and *attended* representations of two arguments and applies several nonlinear transformations as follows:

$$p = f(\left[c_o^{Arg1}; c_a^{Arg1}; c_o^{Arg2}, c_a^{Arg2}\right]) \in \mathbb{R}^{6d_1} \tag{11}$$

, where $f(\cdot) = {}^{\tanh(\cdot)}/_{\|\tanh(\cdot)\|}$. The recognition of discourse relation from p is described in Section 5.

Model Variants In the above description, we assume that the attention distribution over an argument is calculated based on itself. That is, we compare the original argument representation of *Arg1* to its own word embedding matrix in order to obtain the attention distribution over words in this argument. This is also done for *Arg2*. We call

this variant self-attention InAtt, or S-InAtt for short, i.e.,

$$c_a^{Arg1} = \text{Attention}(c_o^{Arg1}, \text{X}^{Arg1}) \tag{12}$$

$$c_a^{Arg2} = \text{Attention}(c_o^{Arg2}, \text{X}^{Arg2}) \tag{13}$$

.

Previous work [6, 15] have found that cross-argument interaction is beneficial for discourse semantic discovery. Following them, we further explore a cross-argument attention strategy. We use the original *Arg1*/*Arg2* argument representation to match the *Arg2*/*Arg1* argument embedding matrix in order to obtain the attention distribution over *Arg2*/*Arg1*. We call this variant cross-attention InAtt, or shortly C-InAtt, i.e.,

$$c_a^{Arg1} = \text{Attention}(c_o^{Arg2}, \text{X}^{Arg1}) \tag{14}$$

$$c_a^{Arg2} = \text{Attention}(c_o^{Arg1}, \text{X}^{Arg2}) \tag{15}$$

.

These two variants are investigated and compared in our experiments.

4.2.2. Outer Attention Model

The outer attention model (OutAtt) is inspired by the cognitive psychology [16, 17]. Even if discourse connectives are not provided, humans can still easily succeed in recognizing the relations of discourse arguments. One reason for this, according to cognitive psychology, would be that humans have a semantic memory in mind, which helps them comprehend word senses and further argument meanings via composition. After understanding what two arguments of a discourse convey, humans can easily interpret the discourse relation of the two arguments. This semantic memory, as discussed by Tulving (1972), refers to general knowledge including "words and other verbal symbols, their meaning and referents, about relations among them, and about rules, formulas, and algorithms for manipulating them". It can be retrieved to help disambiguation and comprehension whenever the barrier of cognition occurs.

Accordingly, the OutAtt model assumes there exists an external semantic memory (a pretrained word embedding matrix) that already encodes some world knowledge, and simulates the human behavior by retrieving relation-relevant knowledge from it to

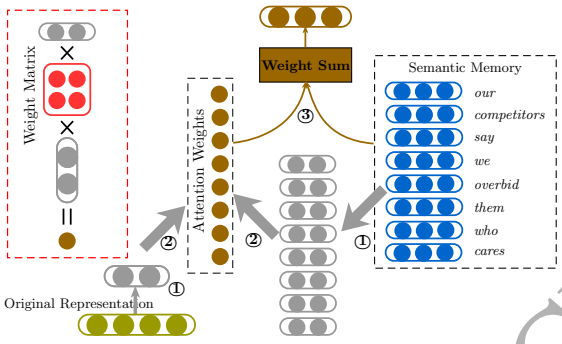


Figure 2: Attention in OutAtt model. We use the shallow and deep yellow color to indicate the original and attended discourse representation respectively. The dashed red box shows the bilinear-style computation for attention weights.

overcome the cognition barrier through an attention network. Formally, the attention mechanism in OutAtt can be summarized as follows:

$$p_a = \text{Attention}(p_o^{Dis}, \text{M}^{Dis}) \tag{16}$$

, where $E = \text{M}^{Dis} \in \mathbb{R}^{d_2 \times m}$ and $v = p_o^{Dis}$ is the semantic memory matrix and original representation for the whole discourse Dis (rather than argument Arg) respectively. d_2 is the dimension of word embedding in the memory, and m is the number of words in the whole discourse. Compared against the word embedding L, our semantic memory M differs significantly in the following two aspects: 1) M is trained on general-topic corpora, while L is discourse-relation-related. This ensures that general knowledge about words can be encoded into M. 2) M is fixed (i.e., not tuned at all) during training, while L must be further tuned to capture relations of discourse arguments. This is because we hope that the semantic and syntactic attributes encoded in the semantic memory can be preserved throughout our neural network. Intuitively, we utilize the task-specific embeddings L to extract related information from general embeddings M so as to enhance the expressiveness of learned discourse representations.

Figure 2 illustrates the architecture of OutAtt. We use the discourse representation in SCNN [7] as our original representation in Eq. (16) due to its simplicity and

effectiveness:

$$p_o^{Dis} = f(\left[c^{Arg1}; c^{Arg2}\right]) \in \mathbb{R}^{6d_1} \tag{17}$$

. As the semantic embedding in M is unchanged throughout our experiment, we employ a bilinear network as our scoring function $g(\cdot)$ to fully model the interaction between discourse and memory(see the read box in Figure 2):

$$g(p_o^{Dis}, \mathrm{M}_i^{Dis}) = p_o^{Dis\prime} W_s \mathrm{M}_i^{Dis\prime} \tag{18}$$

with (as shown by "①" in Figure 2),

$$p_o^{Dis\prime} = \tanh(W_p p_o^{Dis} + b') \tag{19}$$

$$\mathrm{M}_i^{Dis\prime} = \tanh(W_m \mathrm{M}_i^{Dis} + b') \tag{20}$$

, where $p_o^{Dis\prime} \in \mathbb{R}^{d_a}, \mathrm{M}_i^{Dis\prime} \in \mathbb{R}^{d_a}$ are the corresponding representations in a common attention space. In this space, each element in the weight matrix W_s, e.g. $W_{s_{ij}}$, assesses the semantic matching degree of the i-th discourse representation $p_{o_i}^{Dis}$ and the j-th external memory $\mathrm{M}_{i_j}^{Dis}$, which enables the model to learn to automatically strain the discourse with its semantic-related memories. Notice that we differentiate the transformation matrix W_p in Eq. (19) to the W_m in Eq. (20), since the original representation and semantic memory originate from different semantic spaces. However, we share the same bias term for them. This will force our model to learn to encode attention semantics into the transformation matrices, rather than simply the biases.

The OutAtt model represents the whole discourse by concatenating both the original and *attended* discourse representation:

$$p = \left[p_o^{Dis}; p_a\right] \in \mathbb{R}^{6d_1 + d_2} \tag{21}$$

.

4.2.3. Full Attention Model

With InAtt and OutAtt leveraging internal and external information respectively, the full attention model (FullAtt) attempts to combine their strengths together to jointly model both kinds of information. Intuitively, these kinds of information are complementary to each other such that their combination could further boost the recognition performance.

Basically, the FullAtt model is a variant of OutAtt model as formulated in Eq. (16). Nevertheless, instead of using the vanilla discourse representation p_o^{Dis} (see Eq. (17)), FullAtt employs the representation p of InAtt (see Eq. (11)) as its original discourse representation. Therefore, in FullAtt model, the internal information can help provide more evidence to extract more relation-relevant words from the additional semantic memory.

Since there are several variants in InAtt, we further classify our FullAtt model into S-FullAtt and C-FullAtt model accordingly.

5. Parameter Learning

With the learned discourse representation p (see Eq. (11) and Eq. (21)), our models predict the discourse relation using a softmax layer:

$$y = softmax(W_r p + b_r) \tag{22}$$

, where $y \in \mathbb{R}^l$ is the predict discourse relation distribution, and l denotes the number of discourse relations. The $softmax(\cdot)$ function constrains a list of real values into a real-valued distribution.

To assess how well the predicted relation y represents the gold relation y_g, we employ the following cross-entropy error:

$$\mathcal{E}(y, y_g) = -\sum\nolimits_j^l y_{g_j} \times \log\left(y_j\right) \tag{23}$$

. Given a training corpus containing T instances, the joint training objective of our models is to minimize the above error:

$$J(\theta) = \frac{1}{T}\sum\nolimits_{t=1}^{T} \mathcal{E}(y^{(t)}, y_g^{(t)}) + \lambda||\theta||^2 \tag{24}$$

, where λ is the hyperparameter for regularization.

In the phase of parameter initialization, we use the toolkit Word2Vec[1] to pretrain the word embeddings L on a large-scale unlabeled data, and randomly initialize the other

[1] https://code.google.com/p/word2vec/

Relation	# Sentences		
	Train	Dev	Test
COM	1942	197	152
CON	3342	295	279
EXP	7004	671	574
TEM	760	64	85

Table 2: Statistics of implicit discourse relations in the training (Train), development (Dev) and test (Test) sets.

parameters using normal distribution ($\mu = 0, \sigma = 0.01$). To optimize these parameters, we apply the L-BFGS algorithm[2] to the gradient of Eq. (24) which, accordingly, can be computed through standard backpropogation algorithm.

With respect to the semantic memory M^{Dis}, we choose the *GoogleNews-vectors-negative300*[3]. This data contains 300-dimensional vectors (thus, $d_2 = 300$) for 3 million words and phrases. It is trained on part of Google News dataset (about 100 billion words). The wide coverage and newswire domain of its training corpus as well as the syntactic property of Word2Vec models make this vector a good choice for our discussed semantic memory.

6. Experiments

We carried out a series of experiments to evaluate the effectiveness of our models on English implicit DRR task using *PDTB 2.0* corpus[4] [13]. This corpus contains discourse annotations over 2,312 Wall Street Journal articles, and is organized in different sections. Following previous work [14, 18, 19, 7], we used sections 2-20 as training set, sections 21-22 as test set and sections 0-1 as development set for hyperparameter optimization.

We formulated the task in two variants: 1) four separate one-against-all binary

[2] http://www.chokkan.org/software/liblbfgs/

[3] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS 21pQmM/edit?pref=2&pli=1

[4] http://www.seas.upenn.edu/ pdtb/

classification problems: each top-level class of implicit discourse relations vs. the other three discourse relation classes [14, 7, 20]; and 2) four-way classification, which is more natural in realistic settings. With respect to the former, we randomly extracted the same number of positive and negative instances as our training data for each discourse relation class in order to deal with the imbalance issue. However, all instances in the test and development set are kept. With respect to the latter, we employed the instance reweighting trick to deal with the imbalance problem. We reweighted each instance following Rutherford and Xue [21]:

$$w_{ij} = \frac{n}{u_j \cdot k} \tag{25}$$

, where u_j is the total number of instances from class j and k is the number of classes in the dataset of size n. The statistics of various data sets are listed in Table 2.

We employed a large-scale unlabeled data set for word embedding initialization, which contains 1.02M sentences with 33.5M words. We tokenized all data using *Stanford NLP Tool*[5]. According to previous work [7] and preliminary experiments on the development set, we set $d_1 = 128, d_2 = 300, d_a = 64, \lambda = 1e^{-4}$ for all experiments.

6.1. Baseline Methods

We chose two different methods as our baselines: an SVM with feature engineering and a neural network with learned representations:

- **SVM:** a support vector machine (SVM) classifier for relation recognition.[6] We adopted the following features to train SVM: *Bag of Words*, *Cross-Argument Word Pairs*, *Polarity*, *First-Last, First3*, *Production Rules* and *Dependency Rules*. We also used the Brown cluster pair feature [22]. When collecting bag of words, production rules, dependency rules, and cross-argument word pairs, we used a frequency cutoff of 5 for filtering.

- **SCNN:** a shallow convolution neural network for argument representation introduced by Zhang et al. [7].

[5] http://nlp.stanford.edu/software/corenlp.shtml

[6] We used the toolkit *SVM-light*(http://svmlight.joachims .org/) for experiments.

Model	P	R	F1
(R & X [21])	-	-	41.00
(J & E [23])	-	-	35.93
(Chen et al. [24])	-	-	40.17
(L & L [25])	-	-	39.86
SVM	22.79	64.47	33.68
SCNN	22.00	67.76	33.22
S-InAtt	21.80	75.00	33.78
C-InAtt	24.21	60.53	34.59
OutAtt	23.33	61.84	33.87
S-FullAtt	24.35	61.84	34.94
C-FullAtt	23.08	71.05	34.84

(a) COM vs Other

Model	P	R	F1
(R & X [21])	-	-	53.80
(J & E [23])	-	-	52.78
(Chen et al. [24])	-	-	54.76
(L & L [25])	-	-	53.69
SVM	39.14	72.40	50.82
SCNN	39.80	75.29	52.04
S-InAtt	40.55	74.55	52.53
C-InAtt	39.27	77.42	52.11
OutAtt	39.82	80.65	53.32
S-FullAtt	42.36	74.55	54.03
C-FullAtt	41.49	77.78	54.11

(b) CON vs Other

Model	P	R	F1
(R & X [21])	-	-	69.40
(J & E [23])	-	-	80.02
(Chen et al. [24])	-	-	80.62
(L & L [25])	-	-	69.71
SVM	65.89	58.89	62.19
SCNN	56.29	91.11	69.59
S-InAtt	55.11	97.74	70.48
C-InAtt	54.75	99.48	70.62
OutAtt	54.79	99.65	70.70
S-FullAtt	55.22	99.48	71.02
C-FullAtt	55.39	99.30	71.11

(c) EXP vs Other

Model	P	R	F1
(R & X [21])	-	-	33.30
(J & E [23])	-	-	27.63
(Chen et al. [24])	-	-	31.32
(L & L [25])	-	-	37.61
SVM	15.10	68.24	24.73
SCNN	20.22	62.35	30.54
S-InAtt	29.29	34.12	31.52
C-InAtt	28.18	36.47	31.79
OutAtt	23.01	61.18	33.44
S-FullAtt	27.82	43.53	33.94
C-FullAtt	23.83	60.00	34.11

(d) TEM vs Other

Table 3: Recognition results of different models on implicit DRR. **P**=Precision, **R**=Recall, and **F1**=F1 score.

Additionally, we also compared against the published results from six competitive systems:

- **Rutherford and Xue [21]** (R & X [21] for short) convert explicit discourse relations into implicit instances by removing explicit discourse connectives with heuristic rules.
- **Ji and Eisenstein [23]** (J & E [23] for short) explore entity information to augment distributed representations of discourses.
- **Ji et al. [26]** develop a neural language model over sequences of words and treat the discourse relations as latent variables to connect the adjacent sequences.
- **Wu et al. [27]** incorporate synthetic data through bilingual constraint.
- **Chen et al. [24]** propose a deep architecture with gated relevance network for semantic interaction.
- **Liu and Li [25]** (L & L [25] for short) propose to repeatedly read the arguments and dynamically exploit the efficient features useful for recognizing discourse relations.

6.2. Classification Results and Analysis

Our first experiment is for four one-against-all binary classifications. We evaluate each model with several different metrics, including precision, recall, and F1 score. Table 3 summarizes the experiment results, which, overall, show that enhanced with the proposed attention mechanism that highlight contributory discourse words, our models achieve significant performance on all four tasks in terms of F1 score.

We first give an analysis against SCNN, because it is the basis of our model. As shown in Table 3, all our models outperform both SVM and SCNN. Particularly, the C-InAtt/OutAtt/C-FullAtt model gains improvements over SCNN by 1.37%/0.65%/1.62%, 0.07%/1.28%/2.07%, 1.03%/1.11%/1.52% and 1.25%/2.9%/3.57% on COM, CON, EXP and TEM respectively. As the neural baseline, SCNN outperforms SVM on CON, EXP and TEM, but fails on COM. Incorporating our attention networks, however, consistently surpasses SVM and SCNN in all discourse relations. As the only difference

between SCNN and our model lies at the proposed attention mechanism, this result strongly demonstrates that distinguishing different words for relation recognition is indeed helpful.

For InAtt models, we find that C-InAtt almost always yields better results than S-InAtt. We conjecture that the main reason for this lies in cross-argument interactions captured by the cross-argument attention within C-InAtt, which may provide more evidence for discourse relation recognition. Our experiment results on this resonate with the finding of previous work, especially that of Lin et al. [6].

For OutAtt model, we observe that the improvement of OutAtt for relation TEM is the biggest. The gain over SVM is 8.71% and 2.9% over SCNN. As the number of instances in relation TEM is the smallest (see Table 2), we argue that the traditional neural network models may lack of sufficient training instances to capture the underlying discourse relation. However, our OutAtt enhanced with the semantic memory is capable of leveraging the general world knowledge induced from large-scale external corpus to help alleviate this issue.

For FullAtt model, we find that FullAtt obtains slightly better results over both InAtt and OutAtt on all relations. This demonstrates that the internal and external information are complementary to each other, and incorporating them together is beneficial for discourse relation recognition. Additionally, similar as InAtt model, the C-FullAtt model prefers to outperform the S-FullAtt model. This is because the C-InAtt model provides better discourse representation, which further offers much more evidence to extract relation-relevant information from the additional semantic memory.

Compared against InAtt models, OutAtt model tends to achieve better F1 scores. This is reasonable because OutAtt leverages an external semantic memory which is trained on large-scale corpora. The semantic and syntactic attributes of words encoded in this memory is hard be captured by InAtt. Additionally, the improvements of F1 score on COM, CON and TEM mainly result from the improvements of precision, while on EXP it is the recall. The main reason may lie in the quantity difference of training data for different relations, where the training instances of EXP are largest.

For the state-of-the-art results of previous work [23, 21, 24, 25], our models achieve comparable results, even outperforms them on the recognition of some discourse rela-

Model	COM	CON	EXP	TEM	Total
(R & X [21])	44.9/27.6/34.2	49.3/39.6/43.9	61.4/78.8/69.1	38.5/9.1/14.7	57.1/40.5
(Wu et al. [27])	42.1/33.1/37.1	44.2/40.7/42.4	62.6/71.8/66.8	34.5/18.2/23.8	-/42.5
(Ji et al. [26])	-/-/-	-/-/-	-/-/-	-/-/-	59.5/42.3
S-InAtt	28.2/40.8/33.3	45.6/56.3/50.4	65.3/48.6/55.7	23.2/27.1/25.0	47.8/41.1
C-InAtt	31.8/46.7/37.9	45.2/57.0/50.4	66.9/47.6/55.6	23.4/29.4/26.0	48.4/42.5
OutAtt	28.8/39.5/33.3	51.0/45.5/48.1	67.3/57.3/61.9	21.5/36.5/27.1	50.2/42.6
S-FullAtt	33.3/40.1/36.4	48.2/58.1/52.7	69.1/47.6/56.3	22.2/45.9/29.9	49.1/43.8
C-FullAtt	39.8/28.3/33.1	45.8/57.3/51.0	65.0/58.0/61.3	26.4/37.6/31.1	52.1/44.1

Table 4: Performance on the four-way classification task formulation. We show the Precision/Recall/F1 Score for each relation, and provide the Accuracy/Macro-Average F1 score for the whole test set.

tions, e.g. the CON and TEM. This is worthy of efforts given that our models only rely on shallow structure and do not use any manual features designed with prior human knowledge while the previous systems either employ extremely deep and complex network structures or incorporate the neural features together with manual features.

Our second experiment is for four-way classification. For evaluation, we chose the precision/recall/F1 score and accuracy/macro-average F1 score for each relation and the whole test set respectively. The experiment results are shown in Table 4. As in our first experiment, C-InAtt outperforms S-InAtt, OutAtt outperforms InAtt in terms of both accuracy and macro-F1 score on the whole test set. Combining the InAtt and OutAtt into the FullAtt, our model achieves further improvements. Especially, our C-FullAtt yields 52.1% accuracy and 44.1% F1 score on the whole test set respectively, which is competitive against all the previous systems[21, 26, 27]. Although the accuracy of our model is lower than Rutherford and Xue [21] and Ji et al. [26], our model achieves the highest macro-F1 scores on the four-way classification task, a gain of 1.6% in C-FullAtt. This result further demonstrates the superiority of our model.

6.3. Attention Analysis

In order to take a deep look into how the attention mechanism works in our models and whether our models are able to distinguish word contributions, we show one

Model	Relation	Example
C-InAtt	COM	people think of the steel business as an old and mundane smokestack business they 're dead wrong
	EXP	numerous injuries were reported some buildings collapsed , gas and water lines ruptured and fires raged
OutAtt	COM	people think of the steel business as an old and mundane smokestack business they 're dead wrong
	EXP	numerous injuries were reported some buildings collapsed , gas and water lines ruptured and fires raged
C-FullAtt	COM	people think of the steel business as an old and mundane smokestack business they 're dead wrong
	EXP	numerous injuries were reported some buildings collapsed , gas and water lines ruptured and fires raged

Table 5: Examples from the test set. For each argument, some top words with the high attention weights are highlighted in red color.

example for relation COM and EXP from the test set in Table 5, where words assigned with the high attention weights are highlighted in red color.[7]

We find that the attention model indeed learns something that are relevant to the discourse relation recognition task. Our model succeeds in detecting cross-argument phrases that strongly indicate the corresponding relations, e.g., "*think, old, mundane* vs. *wrong*" (COM) and "*injuries* vs. *collapsed, ruptured, raged*" (EXP). However, different models exhibit different capabilities and preferences. Let's consider the example for COM. C-InAtt can recognize that something is "*wrong*". Without external world knowledge, however, C-InAtt can only extract the words "*think, business*" which is hard to answer what is wrong. In this respect, OutAtt succeeds in realizing the "*mundane smokestack*", but fails in detecting this understanding is "*wrong*". Combining their advantages, our C-FullAtt retrieves the words "*think, old, mundane, smokestack, dead, wrong*", which roughly reflects the discourse meaning that *think old mundane smokestack, dead wrong*. Obviously, these words are crucial for discourse comprehen-

[7] We mainly show instances for C-InAtt, OutAtt and C-FullAtt because of their good performance.

sion. All these make the argument representation learned in our model more relation-relevant, and thus boosts the recognition performance.

Additionally, different from manual features *First-Last, First3* defined on fixed positions, our model is able to detect important words on different positions, not limited to the beginning and ending positions. For example, words "*collapsed*", "*ruptured*" with high attentions are almost position-independent, but relation-dependent.

We also notice that not all attentional words are relation-relevant. For example, the word "*and*" (EXP) and the word "*'re*" (TEM) in C-InAtt are rather general. This suggests that the highlighted words here are not necessarily linguistically oriented, since we do not use any linguistic signals for attention training. Instead, our model chooses these words roughly because they are able to benefit the relation recognition.

7. Related Work

Our work is related to previous studies on implicit DRR. It is also in the same line with recent efforts on neural attention mechanisms.

7.1. Implicit Discourse Relation Recognition

The implicit DRR mainly starts from the release of PDTB corpus, a large-scale annotated discourse corpus [13]. Based on this corpus, Pilter et al. [14] perform implicit relation classification using several linguistically informed features. Furthermore, Lin et al. [6] incorporate the context of the two arguments, word pair information, as well as the internal constituent and dependency parses of arguments into their classifier. After that, several more powerful features have been exploited: entities [28], tree kernels [29] and aggregated word pairs [15]. With these features, Park and Cardie [30] performs feature set optimization for better feature combination. Instead of directly classifying discourse relations, predicting appropriate discourse connectives can indirectly help the relation identification [18, 31], while Hong et al. [32] leverage the connective as a bridge to infer the implicit relations from explicit ones. Very recently, Versley [33] explores graph models for the task, and Rutherford and Xue [22] employ brown cluster representations and co-reference patterns.

Most of these methods focus on developing effective hand-crafted features. However, useful features are still likely to be neglected due to the lack of domain knowledge. Instead, we aim to learn these features automatically.

7.2. Neural Network Models

Neural network models have made a great progress in sentence representation learning. Such models include recursive neural networks [34, 35, 36], convolutional neural networks [37] and so on. These models are of great benefit to many downstream NLP applications, such as sentiment classification, question answering, information extraction and machine translation, etc.

In the context of implicit DRR, Braud and Denis [38] investigates the usefulness of unsupervised word representations. Ji and Eisenstein [23] leverage the entity information to enhance the syntactic tree representation, while Zhang et al. [7] exploit a shallow yet effective convolutional neural network with only one hidden layer for argument representation. Following the neural direction, Qin et al. [39] exploit stacked convolutional yet gated neural networks. Ji et al. [26] treat the discourse relation as a latent variable and use neural models to infer it. Chen et al. [24] develop deep neural architecture with a novel gated relevance network to capture semantic interactions between arguments. Liu and Li [25] models the recognition process via repeated reading based on multi-level attention, which, compared with our model, only focus on inner attention. Since we do not focus on the linguistic knowledge, in this work, we follow the research of Zhang et al. [7]. They perform convolution operations on word embeddings. Unfortunately, they equally treat each word in an argument, which, however, might attenuate the effects of relation-sensitive words on the classification of relations. We extend their model and enhance it with a special attention mechanism.

The concerned "attention" have recently gained popularity mainly in multimodal networks, where learning alignments between different modalities is a key interest. For example, Mnih et al. [40] learn image objects and agent actions in the dynamic control problem, and Xu et al. [41] exploit the attention mechanism in the image caption generation task. With respect to neural machine translation, Bahdanau et al. [8] succeed in jointly learning to translate and align words, and Luong et al. [10] further

evaluate different attention architectures on translation. Inspired by these works, we adapt the attention technique to the case of single modal network, and apply it to the implicit DRR.

Additionally, the exploration of semantic memory for implicit DRR is inspired by recent developments in cognitive neuroscience. Yee et al. [17] show how this memory is organized and retrieved in brain. In order to explore semantic memory in neural networks, we borrow ideas from recently introduced memory networks [9, 11, 42, 43] to organize semantic memory as a distributed matrix and use an attention model to retrieve this distributed memory. The adaptation and utilization of semantic memory into implicit DRR, to the best of our knowledge, has never been investigated before.

8. Conclusion and Future Work

In this paper, we have presented two attention-based neural networks for implicit DRR. Instead of assigning each word uniform weights, our model automatically learns attention weights for different words so as to distinguish word contributions to discourse relations. These learned weights can reflect the degrees of importance of corresponding words to discourse relations. Experiment results show that our models are competitive against several strong baselines.

In the future, we would like to exploit different neural network architectures, e.g., deep convolutional neural networks, long-short term memory networks and so on. We are also interested in adapting our model to other similar classification tasks, such as sentiment classification and movie review classification.

Acknowledgments

The authors were supported by National Natural Science Foundation of China (Nos. 61672440, 61622209 and 61403269), Scientific Research Project of National Language Committee of China (Grant No. YB135-49) and Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (No. MJUKF201742).

References

[1] S. Verberne, L. Boves, N. Oostdijk, P.-A. Coppen, Evaluating discourse-based answer extraction for why-question answering, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, ACM, New York, NY, USA, 2007, pp. 735–736. doi:10.1145/1277741.1277883.

[2] P. Cimiano, U. Reyle, J. Šarić, Ontology-driven discourse analysis for information extraction, Data Knowl. Eng. 55 (1) (2005) 59–83. doi:10.1016/j.datak.2004.11.009.

[3] F. Guzmán, S. Joty, L. Màrquez, P. Nakov, Using discourse structure improves machine translation evaluation, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 687–698.

[4] E. Miltsakaki, N. Dinesh, R. Prasad, A. Joshi, B. Webber, Experiments on sense annotations and sense disambiguation of discourse connectives, in: Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT2005), Barcelona, Spain, December, 2005.

[5] E. Pitler, M. Raghupathy, H. Mehta, A. Nenkova, A. Lee, A. Joshi, Easily identifiable discourse relations (2008) 87–90.

[6] Z. Lin, M.-Y. Kan, H. T. Ng, Recognizing implicit discourse relations in the Penn Discourse Treebank, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2009, pp. 343–351.

[7] B. Zhang, J. Su, D. Xiong, Y. Lu, H. Duan, J. Yao, Shallow convolutional neural network for implicit discourse relation recognition, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 2230–2235.

[8] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate.

[9] J. Weston, S. Chopra, A. Bordes, Memory networks, CoRR abs/1410.3916.

[10] T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation (2015) 1412–1421.

[11] S. Sukhbaatar, a. szlam, J. Weston, R. Fergus, End-to-end memory networks, in: C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 2440–2448.

[12] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, R. Socher, Ask me anything: Dynamic memory networks for natural language processing, in: M. F. Balcan, K. Q. Weinberger (Eds.), Proceedings of The 33rd International Conference on Machine Learning, Vol. 48 of Proceedings of Machine Learning Research, PMLR, New York, New York, USA, 2016, pp. 1378–1387.

[13] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, B. Webber, The penn discourse treebank 2.0., in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008, aCL Anthology Identifier: L08-1093.

[14] E. Pitler, A. Louis, A. Nenkova, Automatic sense prediction for implicit discourse relations in text, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, Suntec, Singapore, 2009, pp. 683–691.

[15] O. Biran, K. McKeown, Aggregated word pair features for implicit discourse relation disambiguation, in: Proceedings of the 51st Annual Meeting of the Asso-

ciation for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 69–73.

[16] E. Tulving, Episodic and semantic memory, in: E. Tulving, W. Donaldson (Eds.), Organization of Memory, Academic Press, New York, 1972, pp. 381–403.

[17] E. Yee, E. G. Chrysikou, S. L. Thompson-Schill, The cognitive neuroscience of semantic memory (2014).

[18] Z.-M. Zhou, Y. Xu, Z.-Y. Niu, M. Lan, J. Su, C. L. Tan, Predicting discourse connectives for implicit discourse relation recognition, in: Coling 2010: Posters, Coling 2010 Organizing Committee, Beijing, China, 2010, pp. 1507–1514.

[19] M. Lan, Y. Xu, Z. Niu, Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 476–485.

[20] B. Zhang, D. Xiong, j. su, Q. Liu, R. Ji, H. Duan, M. Zhang, Variational neural discourse relation recognizer, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 382–391.

[21] A. Rutherford, N. Xue, Improving the inference of implicit discourse relations via classifying explicit discourse connectives, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 799–808.

[22] A. Rutherford, N. Xue, Discovering implicit discourse relations through brown cluster pair representation and coreference patterns, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 645–654.

[23] Y. Ji, J. Eisenstein, One vector is not enough: Entity-augmented distributed semantics for discourse relations, Transactions of the Association for Computational Linguistics 3 (2015) 329–344.

[24] J. Chen, Q. Zhang, P. Liu, X. Qiu, X. Huang, Implicit discourse relation detection via a deep architecture with gated relevance network, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1726–1735.

[25] Y. Liu, S. Li, Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1224–1233.

[26] Y. Ji, G. Haffari, J. Eisenstein, A latent variable recurrent neural network for discourse-driven language models, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 332–342.

[27] C. Wu, x. shi, Y. Chen, Y. Huang, j. su, Bilingually-constrained synthetic data for implicit discourse relation recognition, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2306–2312.

[28] A. Louis, A. Joshi, R. Prasad, A. Nenkova, Using entity features to classify implicit discourse relations, in: Proceedings of the SIGDIAL 2010 Conference, Association for Computational Linguistics, Tokyo, Japan, 2010, pp. 59–62.

[29] W. Wang, J. Su, C. L. Tan, Kernel based discourse relation recognition with temporal ordering information, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 710–719.

[30] J. Park, C. Cardie, 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, Association for Computational Linguistics, Seoul, South Korea, 2012, pp. 108–112.

[31] G. Patterson, A. Kehler, Predicting the presence of discourse connectives, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 914–923.

[32] Y. Hong, X. Zhou, T. Che, J. Yao, Q. Zhu, G. Zhou, Cross-argument inference for implicit discourse relation recognition, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12, ACM, New York, NY, USA, 2012, pp. 295–304. doi:10.1145/2396761.2396801.

[33] Y. Versley, Subgraph-based classification of explicit and implicit discourse relations, in: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers, Association for Computational Linguistics, Potsdam, Germany, 2013, pp. 264–275.

[34] R. Socher, C. C.-Y. Lin, A. Y. Ng, C. D. Manning, Parsing natural scenes and natural language with recursive neural networks., in: Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 129–136.

[35] J. Su, D. Xiong, B. Zhang, Y. Liu, J. Yao, M. Zhang, Bilingual correspondence recursive autoencoder for statistical machine translation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1248–1258.

[36] B. Zhang, D. Xiong, J. Su, Battrae: Bidimensional attention-based recursive autoencoders for learning bilingual phrase embeddings., in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), 2017, pp. 3372–3378.

[37] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences (2014) 655–665.

[38] C. Braud, P. Denis, Comparing word representations for implicit discourse relation classification, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 2201–2211.

[39] L. Qin, Z. Zhang, H. Zhao, A stacking gated neural architecture for implicit discourse relation classification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2263–2270.

[40] V. Mnih, N. Heess, A. Graves, k. kavukcuoglu, Recurrent models of visual attention, in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27, Curran Associates, Inc., 2014, pp. 2204–2212.

[41] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.

[42] C. Yan, Y. Zhang, J. Xu, F. Dai, J. Zhang, Q. Dai, F. Wu, Efficient parallel framework for hevc motion estimation on many-core processors, IEEE Transactions on Circuits and Systems for Video Technology 24 (12) (2014) 2077–2089.

[43] C. Yan, Y. Zhang, J. Xu, F. Dai, L. Li, Q. Dai, F. Wu, A highly parallel framework for hevc coding unit partitioning tree decision on many-core processors, IEEE Signal Processing Letters 21 (5) (2014) 573–576.

Biao Zhang received his Bachelor degree in Software Engineering from Xiamen University, and is a graduate student in the School of Software at Xiamen University now. He is supervised by Prof. Hong Duan and Prof. Jinsong Su. His major research interests are natural language processing and deep learning.

Deyi Xiong is a Professor at Soochow University. Previously, he was a Research Scientist at the Institute for Infocomm Research of Singapore from 2007–2013. He completed his Ph.D. in computer science at the Institute of Computing Technology of the Chinese Academy of Sciences in 2007. His research interests are in the area of natural language processing, including parsing and statistical machine translation.

Jinsong Su was born in 1982, he received the Ph.D. degree in Chinese Academy of Sciences. He is now an associate professor of Software School in Xiamen University. His research interests include natural language processing and statistical machine translation.

Min Zhang received his bachelor degree and Ph.D. degree in computer science from Harbin Institute of Technology in 1991 and 1997, respectively. He joined Soochow University in 2013 and is currently a Distinguished Professor with the university. From 1997 to 1999, he was a Postdoctoral Research Fellow with the Korean Advanced Institute of Science and Technology in South Korea. He began his academic and industrial career as a Researcher at Lernout & Hauspie Asia Pacific (Singapore) in 1999. He joined Infotalk Technology (Singapore) as a Researcher in 2001 and became a Senior Research Manager in 2002. He joined the Institute for Infocomm Research (Singapore) in 2003. His current research interests include machine translation, natural language

processing, information extraction, large-scale text processing, and machine learning. He has authored 150 papers in leading journals and conferences. He is the vice president of COLIPS, a steering committee member of PACLIC, an executive member of AFNLP and a member of ACL and IEEE.