SWIM: A Simple Word Interaction Model for Implicit Discourse Relation Recognition

Wenqiang Lei¹, Xuancong Wang², Meichun Liu³, Ilija Ilievski¹, Xiangnan He¹, Min-Yen Kan¹

¹National University of Singapore
²Institute for Infocomm Research
³City University of Hong Kong

wenqiang, xiangnan, knmnyn@comp.nus.edu.sg, wangxc@i2r.a-star.edu.sg
ilija.ilievski@u.nus.edu, meichliu@cityu.edu.hk

Abstract

Capturing the semantic interaction of pairs of words across arguments and proper argument representation are both crucial issues in implicit discourse relation recognition. The current state-of-the-art represents arguments as distributional vectors that are computed via bi-directional Long Short-Term Memory networks (BiLSTMs), known to have significant model complexity.

In contrast, we demonstrate that word-weighted averaging can encode argument representation which can be incorporated with word pair information efficiently. By saving an order of magnitude in parameters and eschewing the recurrent structure, our proposed model achieves equivalent performance, but trains seven times faster.

1 Introduction

Sentences alone do not serve to form coherent discourse. Logical relations, both inter- and intra-sententially, are needed for a coherent text. Such relations are termed discourse relations. Automatically recognizing discourse relations is useful for downstream applications such as machine translation and summarization.

Discourse relations can be overtly signaled by occurrences of explicit discourse connectives such as Indeed and After that (cf Ex. (3) & (4)). In contrast when the context is clear, such overt signals can be omitted, leading to discourse relations that it is said to be implicitly signaled (not marked by a lexical connective in the text; cf Ex. (1) & (2)). The lack of any overt signal makes implicit discourse relations much more challenging to recognize.

This explicit/implicit distinction is adopted by the Penn Discourse Treebank (PDTB, version 2.0) [Prasad et al., 2008]. We adopted the PDTB for this study due to its large size when compared against other discourse corpora such as the Rhetorical Structure Theory Treebank (RST) [Carlson et al., 2002]. While the PDTB has a hierarchical annotation scheme, currently most studies (e.g. [Chandrasekaran et al., 2017]) – inclusive of this one – restrict their use to the top-most, Level–1 categories: Contingency (Ex. (1)), Comparison (Ex. (2)), Expansion (Ex. (3)) and Temporal (Ex. (4)). The two text spans where the discourse relation holds are called arguments, named Arg1 and Arg2 respectively. Arg2 is defined as the argument that syntactically houses the (explicit or implicit) discourse connective.

Modeling word pairs have been shown useful for implicit discourse relation recognition in many studies [Marcu and Echihabi, 2002a; Rutherford and Xue, 2014; Chen et al., 2016]. This is because semantic interactions exist between the two arguments. This can be realised in many forms, of which word pairs are arguably the simplest. For example, the interaction between up and down is most likely to signal a Comparison relation as in Ex. (2). Traditional methods use word pairs [Marcu and Echihabi, 2002a], or variants like Brown Clustering pairs [Rutherford and Xue, 2014], as features for supervised learning.

In addition, argument representation — or how arguments are modeled as a whole — is also crucial for correct interpretation. Taking the fortunate–hurricane pair in Ex. (1), one might construe a Comparison relation due to the word pair's contrasting sentiment polarity. However, it is understood as a Contingency relation when the entire context of both arguments are taken into account.

To address the argument representation challenge, recent works have leveraged the powerful semantic representability

| Ex (1) | You are so fortunate. The hurricane came five hours after you left. |
| Ex (2) | In 1986, the number of cats was down to 1000. In 2002, it went up to 2000. |
| Ex (3) | I never gamble too far. In other words, I quit after one try. |
| Ex (4) | I was sleeping when he entered. |

Figure 1: Toy examples of each of the four Level–1 discourse relations annotated in the PDTB formalism. (1) and (2) are implicit relations; (3) and (4) are explicit. Arg1 is italicized and Arg2 is bolded, as per convention.

*For any questions, please contact Wenqiang Lei and Xuancong Wang.
of neural network models. For example, Gated Relevance Networks (GRN) [Chen et al., 2016] and Neural Networks with Multi-Level Attention (NNMA) [Liu and Li, 2016], apply a bi-directional Long Short-Term Memory network (BiLSTM) [Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997] to represent each argument. Both models achieve the state-of-the-art $F_1$ score without employing handcrafted features. However, BiLSTMs inevitably introduce many parameters. The large number of parameters slows the training process and is prone to overfitting. Especially in the context of the PDTB, a relatively small dataset, a lightweight representation is a viable method to simplify the model.

Motivated by this observation, we propose a new model that integrates the modeling of word pair interaction and argument representation, without the use of BiLSTMs. Our model — termed the Simple Word Interaction Model (SWIM) — achieves a comparable $F_1$ score to the state-of-the-art and runs seven times faster by eschewing the use of BiLSTMs.

2 Related Work

Supervised learning approaches for the implicit discourse relations recognition is the common paradigm in prior work. Surface features for the task include word pairs [Marcu and Echihabi, 2002b] sentiment polarity scores, General Inquirer tags [Pitler et al., 2009], and parser production rules [Lin et al., 2009]. Among all these models, the naïve feature of word pairs [Marcu and Echihabi, 2002a; Rutherford and Xue, 2014] — which is the co-occurrence frequency of a pair of words, one drawn from each of the two arguments — has proven to be extremely efficient.

Another trend is to employ semi-supervised methods. This is due to the limited amount of annotated data and the relative abundance of weakly-labeled data (i.e., explicitly-signaled instances). For example, Lan et al. [2013] explored multitask learning; Hernault et al. [2010] applied feature vector extension; and Rutherford and Xue [2015] selectively added some explicit instances as implicit training data, according to their connectives.

The recent wave of deep learning approaches to NLP problems leverage word embeddings pre-trained on large corpora to achieve significant gains on tasks involving complex semantics. While many architectural designs have been explored (e.g. [Zhang et al., 2016]), the state-of-the-art deep learning methods on our task — the GRN (discussed in detail later) and NNMA — are fundamentally BiLSTM based, involving many parameters, which leads to inefficiency in training and testing.

3 Simple Word Interaction Model (SWIM)

SWIM models both the fine-grained word pair interaction and coarse-grained argument representation. We first introduce the model here and detail its implementation in the following sections. To capture word pair interaction, we calculate an interaction score for each word pair that measures the importance of the interaction between its component words. For argument representation, we apply a weighted average of the component word pair representations. The argument representation thus encapsulates word pair interaction, and is passed to a final a multilayer perceptron layer (MLP) to determine the final discourse relation (cf Figure 2(a)).

### 3.1 Word Interaction Score

Following current best practices, SWIM’s word interaction score captures both linear and quadratic relations between the two word’s embeddings. Formally, let $M$ and $N$ denote the lengths of Argument 1 (hereafter, Arg1) and Argument 2 (Arg2). Further, let $x_i$, $y_j$ denote the pre-trained (row) word embeddings of $i^{th}$ word of Arg1 and $j^{th}$ word of Arg2, separately. Then for each pair of words $x_i$ and $y_j$, SWIM calculates an interaction score as in Eq. (1):

$$ s_{ij} = x_i A y_j^T + B [x_i, y_j] + c_{ij} $$

where $A \in \mathbb{R}^{d \times d}$, $B \in \mathbb{R}^{1 \times 2d}$ and $c_{ij} \in \mathbb{R}$ are trainable parameters, and $[x_i, y_j]$ is a concatenation of the two word embeddings. $x_i A y_j^T$ models the quadratic relation between the two word embeddings, while $B [x_i, y_j]$ captures linear relationships. $c_{ij}$ is the bias term for the final interaction.

At training, these parameters — $A$, $B$ and $c_{ij}$ — model and encode word pair semantics, assigning a high interaction score when well correlated with a particular discourse relation class. Similar approaches have been applied to many fields, e.g. recommendation system [He et al., 2017].

SWIM calculates the interaction matrix $S$, computing $s_{ij}$ for each possible ($i^{th}$) word of Arg1 and ($j^{th}$) word of Arg2.

### 3.2 Argument Representation

Both (Bi)LSTMs and embedding averaging are valid methods for representing text sequences, inclusive of discourse arguments and sentences. Recent studies suggest that LSTMs perform well when sequential order is important [Iyyer et al., 2015]. However, this is less the case in argument representation, where content plays a larger role than ordering — e.g., in Exs. (1) & (2), the discourse relations do not change even when the words are reordered. Simpler methods, such word averaging may be sufficient and effective as suggested by Wieting et al. [2016] who concluded that “word averaging models perform well for [the related tasks of] sentence similarity and entailment, outperforming LSTMs.” Hence, SWIM adopts embedding averaging.

However, embedding averaging alone is insufficient – each argument’s words are actually understood in the context of the opposing argument. For example, on its own down in Arg1 of Ex. (2) is less likely considered to signal any discourse relation. However, once combined with the word up in Arg2, it becomes the most important signal in Arg1 for
the Comparison relation. In light of this, SWIM represents each argument as an average of its component words’ word embeddings, weighted for its interaction with the opposing argument.

We denote SWIM’s argument representation for Arg1 (Arg2) as \( x' (y') \), as calculated in Eq. (2).

\[
x' = \frac{1}{M} \sum_{i=1}^{M} (\sum_{j=1}^{N} \frac{exp(s_{ij})}{\sum_{k=1}^{M} exp(s_{kj})}) [x_i, y_j] \\
y' = \frac{1}{N} \sum_{j=1}^{N} (\sum_{i=1}^{M} \frac{exp(s_{ij})}{\sum_{k=1}^{M} exp(s_{kj})}) [y_j, x_i] \tag{2}
\]

In Eq. (2), we start with the concatenations of word pair embeddings \([x_i, y_j]\), weighting them with the interaction score \( s_{ij} \), to account for its importance.

Let us walk through Eq. (2), taking the computation of \( x' \) as an example. For each word \( x_i \) in Arg1, we enumerate all words in Arg2 \( y_j \) for \( j \in \{1, 2...N\} \) to form word pairs \( ([x_i, y_j]) \). We weight these word pair representation according to its normalized interaction score, obtaining an interaction weighted word representation, the term in the parentheses. All \( M \) interaction-enhanced word representations are averaged to arrive at the final form for \( x' \).

SWIM computes a single representation of both Arg1 and Arg2 by concatenating \( x' \) and \( y' \) for input to the final MLP classification to obtain the output discourse relation:

\[
output = f_o(W_o f_h(W_h [x', y'])) \tag{3}
\]

where \( W_h \in \mathbb{R}^{d \times k} \) and \( W_o \in \mathbb{R}^{k \times n} \). In \( W_o \) and \( W_h \), \( n \) is the number of class labels, \( d \) is the embedding size, \( k \) is the size of the hidden layer. \( f_h \) and \( f_o \) are sigmoid activation functions. The final classification layer described in Eq. (3) is a two-layer MLP whose hidden layer is designed to obtain a more abstract representation.

### 3.3 Model Discussion On BiLSTMs

Both state-of-the-art prior models for implicit discourse relation classification – GRN and NNMA – adopt BiLSTMs for argument representation. A discussion on the role of BiLSTMs in both models is relevant. We use GRN as a sample for this discussion as i) both GRN and our proposed SWIM is designed to model word pair interaction and apply similar word interaction calculation approach; ii) the necessity of BiLSTMs in GRN has been previously studied in [Chen et al., 2016].

The GRN workflow is illustrated in Figure 2(b). GRN feeds one argument into an individual BiLSTMs to get an intermediate representation for each word, denoted as \( h_{x_i} \) and \( h_{y_j} \), for each \( x_i \) and \( y_j \), separately. As illustrated in Figure 3, the intermediate representation incorporates information from the argument as a whole.

For each pair of \( h_{x_i} \) and \( h_{y_j} \), an interaction score \( s_{ij} \) is calculated as in Eqs. (4) & (5). As we focus on the role of BiLSTMs, we do not discuss interaction scores further.

\[
s_{ij} = u(g \odot h_{x_i} M^{[1:r]} h_{y_j}^T + (1 - g) \odot f(V[h_{x_i}, h_{y_j}] + b) \tag{4}
\]

\[
g = \sigma(W_o [h_{x_i}, h_{y_j}]) \tag{5}
\]

In the above, \( \odot \) denotes element-wise multiplication.

### 4 Experiments

We now evaluate SWIM, with the aim to show its competitiveness in terms of prediction performance. Importantly, we note that the evaluation of a computational model not only...
lies in the best prediction performance, but also its stability (robustness), and efficiency. This is especially important for neural network models, which are computationally intensive and sensitive to random factors. We first describe the systems we compare against, then describe the experimental setup before describing the main results.

### 4.1 Comparison Systems

In the following list of architectures, numberings follow those used in Table 2 for convenience. First, we must validate performance against the current state-of-the-art baselines in terms of raw prediction performance, in the numbered systems below. The cited papers, our SWIM and replicated experiments follows the same training, development, and test-splits, so are comparable.

1-2. Gated Relevance Networks (GRN; [Chen et al., 2016]). We replicate their architecture with the assistance of the original author, tuning parameters according to the original paper.

3-4. Neural Networks with Multi-Level Attention (NNMA; [Liu and Li, 2016]). We cite results from the paper.

We study a number of variants of SWIM to test our design decisions. In the lettered systems below, we examine variations on argument representation, using embedding averaging and BiLSTMs:

a. Embedding Average (EA). We perform naïve embedding averaging to get a representation vector for each argument. Then, we concatenate the two vectors for input to a MLP as in Eq. (3).

b. BiLSTM. We use the word embeddings as input to BiLSTMs $(h^f_t, h^b_t)$, concatenate their output to form the argument representation and pass it to the MLP layer for classification. This setting benchmarks the vanilla use of BiLSTM for argument representation.

c. Word Interaction Score (WIS). We use scores in the interaction matrix $S$ for classification without modification. Following the GRN approach (cf Section 3.3), we perform max-pooling on $S$ before inputting the scores to the MLP layer. This tests whether our weighted argument representation helps.

d. BiLSTMs + WIS. Following GRN, we feed word embeddings to BiLSTMs to get an intermediate representation. Then we calculate word interaction scores based on these intermediate representation using Eq. (1), obtaining the interaction matrix $S$ which is max-pooled and fed into a MLP layer. This aims to test the effect of BiLSTMs on WIS. BiLSTM + WIS is similar to GRN except for the word interaction computation part. However, we here only focus on the study of BiLSTMs without discussion its difference.

e. BiLSTM + SWIM. As described in Section 3.3, this tests whether BiLSTMs further aid SWIM, since content information has already been encoded by SWIM’s word embedding averaging.

We also investigate whether both linear and quadratic terms are necessary for SWIM’s word interaction computation:

A. Quadratic Term (QT only). Only the quadratic term and bias term of Eq. (1) are retained.

B. Linear Term (LT only). Only the linear and bias terms are retained.

### 4.2 Experimental Setting

We adopt the standard settings for the PDTB v2.0 dataset used in our experimentation (Sections 2–20, Sections 0-1 & Sections 23-24, and Sections 21-22 for training, development and testing, respectively.

We follow the practice of [Zhou et al., 2010; Liu and Li, 2016] which follow the standard definition of PDTB v2.0, which admits a separate categorization outside of discourse relations for EntRel (hence distinct from Expansion), unlike [Pitler et al., 2009; Chen et al., 2016] that deemed entity relations (EntRel) as a form of Expansion relations. Aside from this, our other settings are standard: we use Stanford CoreNLP [Manning et al., 2014] for tokenization, pad all sentences to length 50, and use Stanford’s GloVe [Pennington et al., 2014] 100 dimensional pre-trained word embeddings for SWIM and 50 dimensional pre-trained embedding for BiLSTMs + SWIM. The embedding layer is fixed during training, and dropout is performed on the input and MLP layers (dropout percentage = 20%). For training, we adopt multi-class cross-entropy loss, using AdaGrad for the stochastic optimization [Duchi et al., 2011]. The initial learning rate is set at 0.01, with a batch size of 32. Following [Liu and Li, 2016; Rutherford and Xue, 2014], we use instance re-weighting.

### 4.3 Experimental Results

Table 2 presents our results. Prediction performance is evaluated in both standard schemes of $n$ binary classifications and a single $n$-way classification (where $n = 4$), using macro $F_1$.

The upper portion of Table 2 shows that our SWIM architecture (Row 5) performs on par with GRN and NNMA, with the exception of Temporal relations. Introspecting these results at an instance level, we find that GRN, NNMA and SWIM exhibit little agreement on Temporal relations. As Temporal relations only constitute 5% of the dataset, it is a minority class where its performance may be largely affected by random factors. In our efficiency analysis, we calculate

<table>
<thead>
<tr>
<th>Model</th>
<th>$d = 50$</th>
<th>$d = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. BiLSTM</td>
<td>80k</td>
<td>320k</td>
</tr>
<tr>
<td>2. GRN</td>
<td>100k</td>
<td>400k</td>
</tr>
<tr>
<td>3. NNMA (2 levels)</td>
<td>260k</td>
<td>720k</td>
</tr>
<tr>
<td>4. NNMA (3 levels)</td>
<td>350k</td>
<td>920k</td>
</tr>
<tr>
<td>5. SWIM</td>
<td>12.5k</td>
<td>30k</td>
</tr>
<tr>
<td>6. BiLSTM + SWIM</td>
<td>102.5k</td>
<td>390k</td>
</tr>
</tbody>
</table>

Table 1: Model complexity analysis. Number of parameters versus the hidden size $d$. We omit matrices with fewer than 1k parameters.
the time cost per instance in one training epoch (measured in milliseconds). Our proposed SWIM runs 7 times (3.46/0.49) faster than GRN.

This result is significant as the training times listed in Table 2 are for individual epochs and for a single architecture. The time savings benefits two aspects: first, neural models are typically trained over hundreds of epochs and with hundreds of individual configurations for hyperparameter tuning, leading to significant savings in development and tuning times in creating the final model. Second, semi-supervised learning is an important direction for implicit discourse relation recognition, due to the abundance of weakly labeled data. In semi-supervised learning, usually a much large number of training instances are involved. For example, Liu et al. [2016] report supervised learning, usually a much large number of training instances are involved. For example, Liu et al. [2016] report supervised learning, usually a much large number of training instances are involved. For example, Liu et al. [2016] report supervised learning, usually a much large number of training instances are involved. For example, Liu et al. [2016] report supervised learning, usually a much large number of training instances are involved. For example, Liu et al. [2016] report.

Table 2: Models' effectiveness, efficiency and stability. Effectiveness measured by macro $F_1$; efficiency, by average processing time per instance, in milliseconds; stability, by macro-averaged standard deviation over 10 runs.

<table>
<thead>
<tr>
<th>Model</th>
<th>Comp.</th>
<th>Cnt.</th>
<th>Exp.</th>
<th>Temp.</th>
<th>4-way</th>
<th>Time (ms)</th>
<th>Avg.Std.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GRN (cited)</td>
<td>40.17%</td>
<td>54.76%</td>
<td>–</td>
<td>31.32%</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2. GRN (replicated)</td>
<td>39.05%</td>
<td>54.53%</td>
<td>69.01%</td>
<td>33.52%</td>
<td>44.61%</td>
<td>3.46</td>
<td>0.012</td>
</tr>
<tr>
<td>3. NNMA (cited, two level)</td>
<td>39.86%</td>
<td>53.69%</td>
<td>69.71%</td>
<td>37.61%</td>
<td>44.95%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>4. NNMA (cited, three level)</td>
<td>36.70%</td>
<td>54.48%</td>
<td><strong>70.43%</strong></td>
<td><strong>38.84%</strong></td>
<td>46.25%</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>5. SWIM</td>
<td><strong>40.47%</strong></td>
<td><strong>55.36%</strong></td>
<td>69.5%</td>
<td>35.34%</td>
<td><strong>46.46%</strong></td>
<td>0.49</td>
<td>0.008</td>
</tr>
<tr>
<td>a. EA</td>
<td>33.01%</td>
<td>46.56%</td>
<td>68.31%</td>
<td>29.59%</td>
<td>40.22%</td>
<td>0.08</td>
<td>0.005</td>
</tr>
<tr>
<td>b. BiLSTM</td>
<td>34.01%</td>
<td>47.31%</td>
<td>68.53%</td>
<td>30.01%</td>
<td>40.84%</td>
<td>3.0</td>
<td>0.011</td>
</tr>
<tr>
<td>c. WIS</td>
<td>35.53%</td>
<td>51.15%</td>
<td>68.98%</td>
<td>30.72%</td>
<td>40.64%</td>
<td>0.45</td>
<td>0.007</td>
</tr>
<tr>
<td>d. BiLSTM + WIS</td>
<td>39.42%</td>
<td>53.85%</td>
<td>70.05%</td>
<td>32.75%</td>
<td>44.57%</td>
<td>3.37</td>
<td>0.011</td>
</tr>
<tr>
<td>e. BiLSTM + SWIM</td>
<td>38.71%</td>
<td>54.32%</td>
<td>70.02%</td>
<td>35.06%</td>
<td>45.96%</td>
<td>3.37</td>
<td>0.012</td>
</tr>
<tr>
<td>A. QT only</td>
<td>37.73%</td>
<td>53.02%</td>
<td>68.12%</td>
<td>33.21%</td>
<td>43.25%</td>
<td>0.34</td>
<td>0.008</td>
</tr>
<tr>
<td>B. LT only</td>
<td>36.78%</td>
<td>51.51%</td>
<td>67.41%</td>
<td>32.72%</td>
<td>42.04%</td>
<td>0.18</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Table 3: Word order sensitivity comparison: $F_1$ of a single 4-way classification on the scrambled ($test'$) set.

<table>
<thead>
<tr>
<th></th>
<th>GRN</th>
<th>SWIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$test'$</td>
<td>42.35% (−2.26%)</td>
<td>46.46% (NC)</td>
</tr>
</tbody>
</table>

STM may contribute noise over the use of the content cues already encoded by embedding averaged argument representation. Additionally, all BiLSTMs models take significantly longer training times and produce less stable results (cf Row 5 vs. Rows 2, b, d, and e).

Finally, comparing SWIM against the selective use of only the quadratic or linear word interaction terms in the bottom portion of the table suggest that both are needed to properly model word interaction (Row 5 vs. Rows A and B).

Overall, we conclude that the current SWIM configuration strikes a good balance among effectiveness, stability, efficiency and model complexity.

Discussion

There are loose ends to validate in our claims that have not yet been supported by evidence. Here, we detail three issues where we present auxiliary data to buttress our arguments.

First, we have argued that positional information encoded by BiLSTM (in the guise of GRNs) is not needed for our classification task. We validate this proposition via supplemental experimentation. To give credence that GRN models positional information, we randomize the order of the words in each argument in the test set to obtain a new test set ($test'$). We then test the pre-trained GRN, SWIM against $test'$. Table 3 gives the 4-way classification performance figures on $test'$. The results indicate that GRN is sensitive to the position of words, while SWIM-based models retain their performance, validating that SWIM does not model positional information.

Second, we need to assess the sensitivity of the neural architectures to dataset scale, as we have argued that the SWIM architecture is optimal for the task, independent of the scale of data currently available. We train and test GRN and SWIM with different model sizes by changing the hidden layer and
embedding sizes\(^3\) of each model to have their total number of parameters be at parity: set to 10k, 30k, 100k and 300k (Table 4). This result suggests that GRN intrinsically requires much more parameters to achieve better prediction results; in contrast, SWIM makes better use of its limited number of parameters.

<table>
<thead>
<tr>
<th>Model</th>
<th>10k</th>
<th>30k</th>
<th>100k</th>
<th>300k</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRN</td>
<td>42.06%</td>
<td>43.03%</td>
<td><strong>44.61%</strong></td>
<td>43.59%</td>
</tr>
<tr>
<td>SWIM</td>
<td>45.93%</td>
<td><strong>46.46%</strong></td>
<td>44.32%</td>
<td>43.95%</td>
</tr>
</tbody>
</table>

Table 4: \(F_1\) score of 4-way classification with different total number of parameters. The optimal configuration is reported in Table 2, and bolded per system.

Finally, we also wish to assess whether SWIM actually does assigns high interaction scores to important word pairs in the matrix \(S\) through per instance analysis. In Figure 4, we introspect the interaction matrices of WIS and SWIM on a PDTB example to illustrate how they differently they behave. We note that both WIS and SWIM calculate word interaction scores in the same way, but that SWIM’s contextualization of its argument representation modeling assigns high scores more intuitively than WIS does. WIS assigns word pairs featuring are (‘re) high scores, possibly due to its cor-

\(\text{pus frequency. The arithmetic post-processing of the argument context by SWIM redistributes scores, correctly assigning the fortunate–hurricane pair the highest interaction score. Separately, Figure 5 shows SWIM’s interaction scores on a Comparison instance, where the words (row) down and (column) up are assigned high weights, which is in accord with our intuition that the word pair is critical in determining the discourse relation. Interestingly, the word (row) about also yields a high score. What was the cause for this? Drilling down, we find that the Wall Street Journal uses similar textual expressions “[numbers] ... increase/decrease/up/down ... about” when reporting many instances of financial news. The high weight for about lends evidence that SWIM can discover such important contextual co-occurrence corpus patterns.}

5 Conclusion
We proposed a simple neural model for implicit discourse relation recognition, named SWIM, which accounts for both word pair interaction and argument representation. In contrast to previous works, we utilize word embedding average, instead of BiLSTMs, for argument representation. Experiment results show our model is more stable and runs faster while still achieving state-of-the-art \(F_1\) scores.

In the wider context of neural network research, our work finds additional evidence that BiLSTMs spend much effort to model positional information, which we have shown to be less helpful for our task. In tasks where word content is valued over word ordering information, our work suggests that simpler models such as an embedding averaging, can replace BiLSTMs while achieving similar performance.

Acknowledgments
We gratefully acknowledge Qian Zhong of the Department of Linguistics and Translation, City University of Hong Kong for fruitful discussions on this research, and also acknowledge the support of NVIDIA Corporation for their donation of the Titan X GPU that facilitated this research.

\(^3\)For GRN, 50d is the smallest pre-trained GloVe word embedding. Therefore, we have to reduce the hidden size of GRN. For SWIM, we use larger pre-trained embeddings and enlarge the hidden size to increase model size.
References


