Discourse Relation Prediction: Revisiting Word Pairs with Convolutional Networks

Siddharth Varia
Christopher Hidey
Tuhin Chakrabarty

Discourse Relation Prediction

Penn Discourse Tree Bank (PDTB) - shallow discourse semantics between segments

- Classes
  - Comparison
  - Expansion
  - Contingency
  - Temporal
- Relation Types
  - Explicit
  - Implicit

Discourse Relation Prediction

Penn Discourse Tree Bank (PDTB) - shallow discourse semantics between segments

- Classes
  - Comparison
  - Expansion
  - Contingency
  - Temporal
- Relation Types
  - Explicit
  - Implicit

Implicit Example:

Arg. 1: Mr. Hahn began selling non-core businesses, such as oil and gas and chemicals.

Arg 2. He even sold one unit that made vinyl checkbook covers.

Discourse Relation Prediction

Penn Discourse Tree Bank (PDTB) - shallow discourse semantics between segments

- Classes
  - Comparison
  - Expansion
  - Contingency
  - Temporal
- Relation Types
  - Explicit
  - Implicit

Implicit Example:

Arg. 1: Mr. Hahn began selling non-core businesses, such as oil and gas and chemicals.

[Expansion/***in fact***]

Arg 2. He even sold one unit that made vinyl checkbook covers.

Outline

- Background
- Related Work
- Method
- Results
- Analysis and Conclusions

Background

John is good in math and sciences.

Paul fails almost every class he takes.

Daniel Marcu and Abdessamad Echihabi. An Unsupervised Approach to Recognizing Discourse Relations. ACL 2002.

Background

John is good in math and sciences.

Paul fails almost every class he takes.

[COMPARISON]

Daniel Marcu and Abdessamad Echihabi. An Unsupervised Approach to Recognizing Discourse Relations. ACL 2002.

Background

John is **good** in math and sciences.

Paul **fails** almost every class he takes.

[**COMPARISON**]

Daniel Marcu and Abdessamad Echihabi. An Unsupervised Approach to Recognizing Discourse Relations. ACL 2002.

Related work

- Word Pairs
  - Cross-product of words on either side of the connective (Marcu and Echihabi, 2002; Blair-Goldensohn et al., 2007)
  - Top word pairs are discourse connectives and functional words (Pitler, 2009)
  - Separate TF-IDF word pair features for each connective (Biran and McKeown, 2013)

Related work

- Word Pairs
  - Cross-product of words on either side of the connective (Marcu and Echihabi, 2002; Blair-Goldensohn et al., 2007)
  - Top word pairs are discourse connectives and functional words (Pitler, 2009)
  - Separate TF-IDF word pair features for each connective (Biran and McKeown, 2013)
- Neural Models
  - Jointly modeling PDTB and other corpora (Liu et al., 2016; Lan et al., 2017)
  - Adversarial learning of model with connective and model without (Qin et al., 2017)
  - Jointly modeling explicit and implicit relations using full paragraph context (Dai and Huang, 2018)

Research Questions

1. Can we explicitly model word pairs using neural models?

2. Can we transfer knowledge from labeled explicit examples in the PDTB?

Method

I am late for the meeting **because** the train was delayed.

Method

Arg. 2

Arg. 1

	because	the	train	was	delayed	.
I						
am						
late						
for						
the						
meeting						

Method

Arg. 2

Arg. 1

	because	the	train	was	delayed
late	late,because	late,the	late,train	late,was	late,delayed
for	for,because	for,the	for,train	for,was	for,delayed
the	the,because	the,the	the,train	the,was	the,delayed
meeting	meeting, because	meeting, the	meeting, train	meeting, was	meeting, delayed

Arg. 1 x Arg. 2

Method

Arg. 2

Arg. 1

	[redacted]	the	train	was	delayed
late	[redacted]	late,the	late,train	late,was	late,delayed
for	[redacted]	for,the	for,train	for,was	for,delayed
the	[redacted]	the,the	the,train	the,was	the,delayed
meeting	[redacted]	meeting, the	meeting, train	meeting, was	meeting, delayed

Same for implicit, minus connective

Arg. 1 x Arg. 2

Arg 1: I was [late] for the meeting

Arg 2: [because] the train was delayed.

Method

late	because	late	the	late	train	late	was	late	delayed
for	because	for	the	for	train	for	was	for	delayed
the	because	the	the	the	train	the	was	the	delayed
meeting	because	meeting	the	meeting	train	meeting	was	meeting	delayed

Convolutions over Word/Word Pairs (WP-1)

Method

Arg 1: I was **[late]** for the meeting

Arg 2: because **[the]** train was delayed.

late	because	**late**	**the**	late	train	late	was	late	delayed
for	because	for	the	for	train	for	was	for	delayed
the	because	the	the	the	train	the	was	the	delayed
meeting	because	meeting	the	meeting	train	meeting	was	meeting	delayed

Convolutions over Word/Word Pairs (WP-1)

Method

Arg 1: I was **[late]** for the meeting

Arg 2: because the **[train]** was delayed.

late	because	late	the	**late**	**train**	late	was	late	delayed
for	because	for	the	for	train	for	was	for	delayed
the	because	the	the	the	train	the	was	the	delayed
meeting	because	meeting	the	meeting	train	meeting	was	meeting	delayed

Convolutions over Word/Word Pairs (WP-1)

Method

Arg 1: I was **[late]** for the meeting

Arg 2: because the train **[was]** delayed.

late	because	late	the	late	train	**late**	**was**	late	delayed
for	because	for	the	for	train	for	was	for	delayed
the	because	the	the	the	train	the	was	the	delayed
meeting	because	meeting	the	meeting	train	meeting	was	meeting	delayed

Convolutions over Word/Word Pairs (WP-1)

Method

Arg 1: I was **[late]** for the meeting

Arg 2: [because the train was] delayed.

late	**because**	**late**	**the**	**late**	**train**	**late**	**was**	late	delayed
for	because	for	the	for	train	for	was	for	delayed
the	because	the	the	the	train	the	was	the	delayed
meeting	because	meeting	the	meeting	train	meeting	was	meeting	delayed

Convolutions over Word/N-gram Pairs (WP-N)

Method

Arg 1: I was **[late]** for the meeting

Arg 2: because **[the train was delayed]**.

late	because	**late**	**the**	**late**	**train**	**late**	**was**	**late**	**delayed**
for	because	for	the	for	train	for	was	for	delayed
the	because	the	the	the	train	the	was	the	delayed
meeting	because	meeting	the	meeting	train	meeting	was	meeting	delayed

Convolutions over Word/N-gram Pairs (WP-N)

Method

Arg 1: I was **[late]** **[for]** the meeting

Arg 2: **[because]** the **[train was delayed]**.

late	because	late	the	**late**	**train**	**late**	**was**	**late**	**delayed**
for	**because**	for	the	for	train	for	was	for	delayed
the	because	the	the	the	train	the	was	the	delayed
meeting	because	meeting	the	meeting	train	meeting	was	meeting	delayed

Convolutions over Word/N-gram Pairs (WP-N)

Method

Arg 1: I was **[late]** **[for]** the meeting

Arg 2: **[because the]** train **[was delayed]**.

late	because	late	the	late	train	**late**	**was**	**late**	**delayed**
for	**because**	**for**	**the**	for	train	for	was	for	delayed
the	because	the	the	the	train	the	was	the	delayed
meeting	because	meeting	the	meeting	train	meeting	was	meeting	delayed

Convolutions over Word/N-gram Pairs (WP-N)

Method

Arg 1: I was **[late for the meeting]**

Arg 2: [because] the train was delayed.

late	**because**	late	the	late	train	late	was	late	delayed
for	**because**	for	the	for	train	for	was	for	delayed
the	**because**	the	the	the	train	the	was	the	delayed
meeting	**because**	meeting	the	meeting	train	meeting	was	meeting	delayed

Convolutions over Word/N-gram Pairs (WP-N)

Method

Arg 1: I was **[late] [for the meeting]**

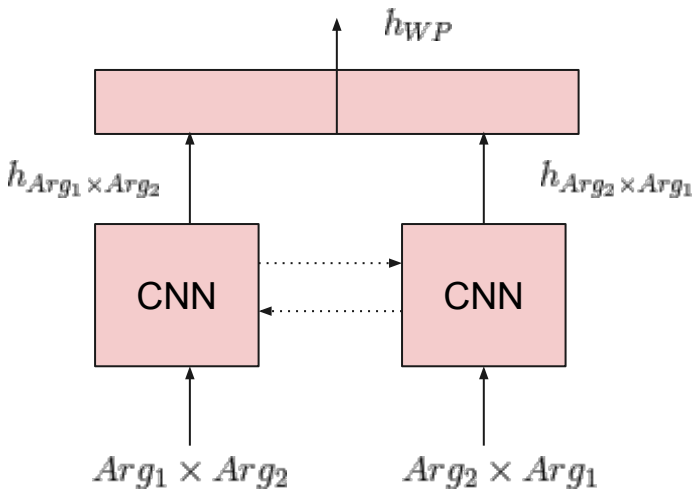
Arg 2: **[because] [the]** train was delayed.

late	because	late	the	late	train	late	was	late	delayed
for	**because**	for	the	for	train	for	was	for	delayed
the	**because**	the	the	the	train	the	was	the	delayed
meeting	**because**	meeting	the	meeting	train	meeting	was	meeting	delayed

Convolutions over Word/N-gram Pairs (WP-N)

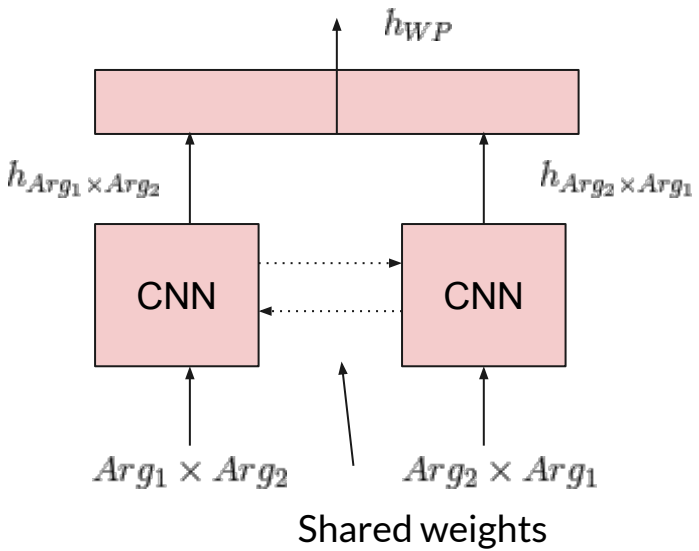
Method

Word/Word and Word/N-gram Pairs (WP-N)

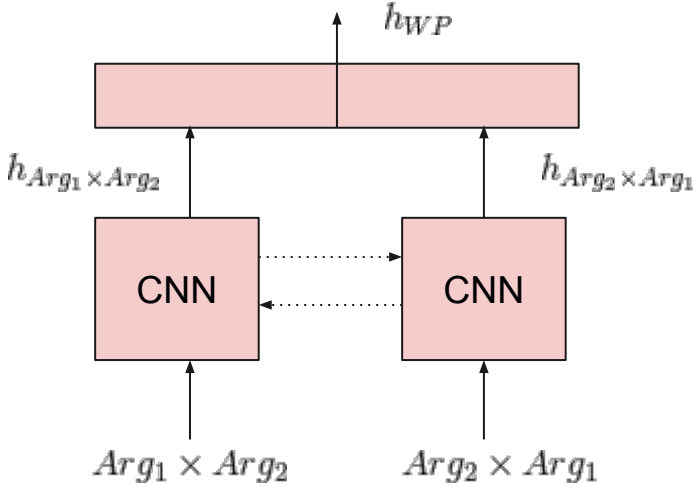


Method

Word/Word and Word/N-gram Pairs (WP-N)

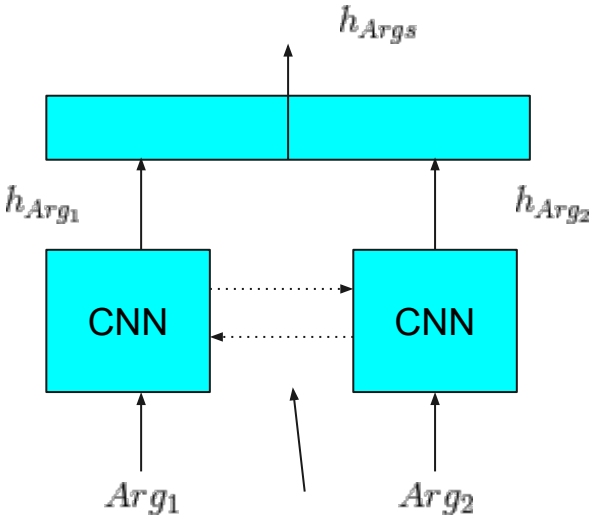


Method



Word/Word and Word/N-gram Pairs (WP-N)

Individual Arguments

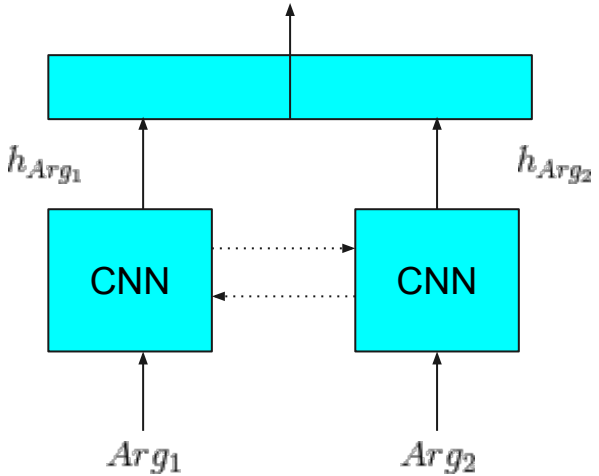


Shared weights

Method

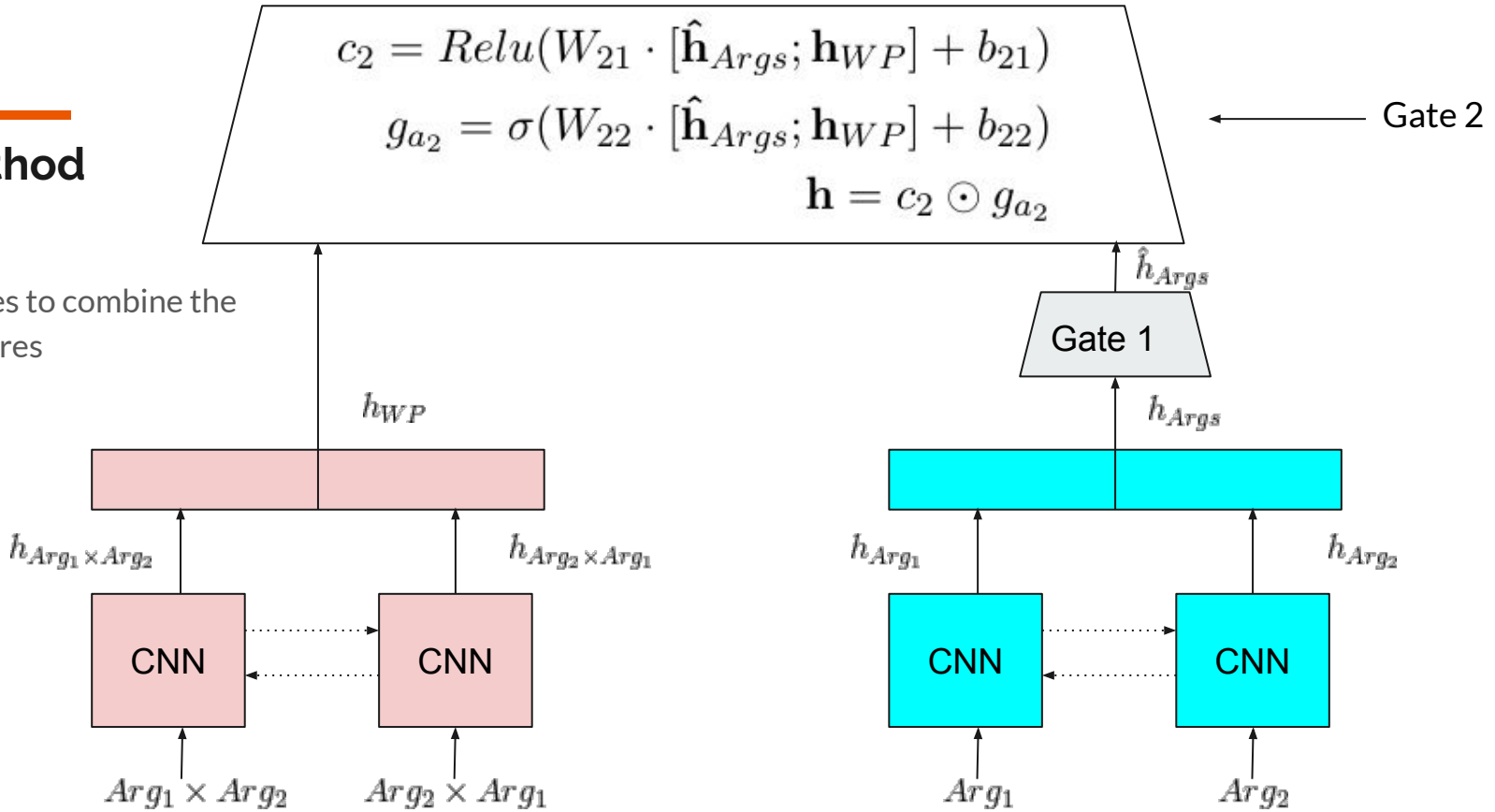
$$c_1 = Relu(W_{11} \cdot [\mathbf{h}_{Arg1}; \mathbf{h}_{Arg2}] + b_{11})$$
$$g_{a_1} = \sigma(W_{12} \cdot [\mathbf{h}_{Arg1}; \mathbf{h}_{Arg2}] + b_{12})$$
$$\hat{\mathbf{h}}_{Args} = c_1 \odot g_{a_1}$$

Gate 1



Method

Identical gates to combine the various features

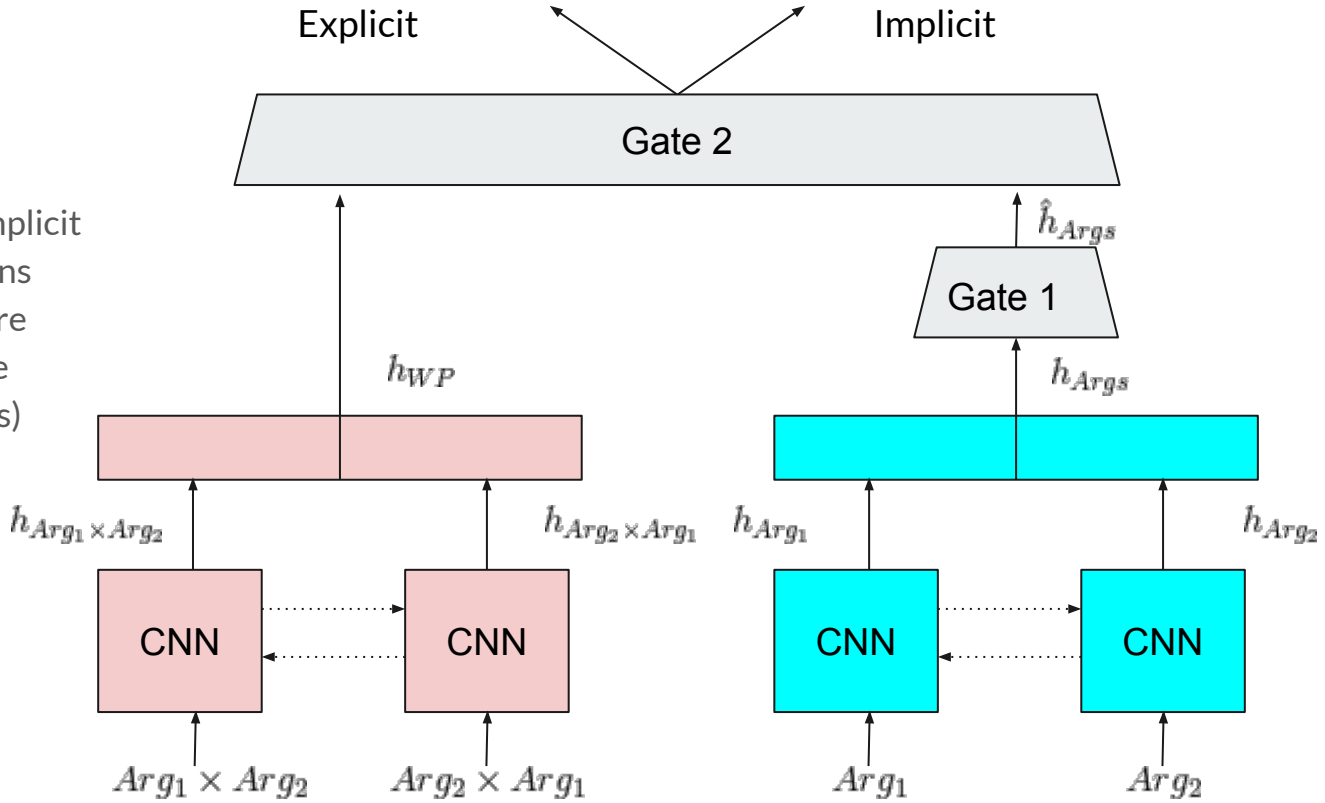


Word/Word and Word/N-gram Pairs (WP-N)

Individual Arguments

Method

Joint learning of implicit and explicit relations (shared architecture except for separate classification layers)

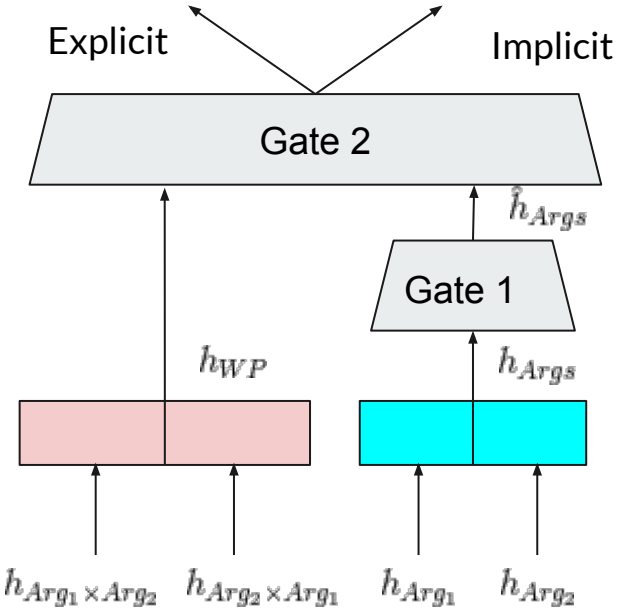


Word/Word and Word/N-gram Pairs (WP-N)

Individual Arguments

Experimental Settings

- Features from Arg. 1 and Arg. 2:
  - **Word/Word Pairs**
  - **Word/N-Gram Pairs**
  - **N-gram features**
- WP - filters of sizes 2, 4, 6, 8
- N-gram - filters of sizes 2, 3, 4, 5
- Static word embeddings and one-hot POS encoding



Dataset and Experiments

- We evaluate our architecture on two different datasets:
  - PDTB 2.0 (for binary and four-way tasks)
  - CoNLL 2016 shared task blind test sets (for fifteen-way task)
- We perform evaluation across three different tasks:
  - Binary classification (One vs. All)
  - Four-way classification
  - Fifteen-way classification
- We use the standard train/validation/test splits for the above datasets in line with the previous work for fair comparison

Results on Four-way Task

Results* on Implicit Relations

Model	Macro-F1	Accuracy
Lan et al., 2017	47.80	57.39
Dai & Huang, 2018	(48.82)	(58.2)
Bai & Zhao, 2018	51.06	-
WP-[1-4], Args, Joint Learning	**(50.2)**	**(59.13)**
	51.84	**60.52**

Results* on Explicit Relations

Model	Macro-F1	Accuracy
Dai & Huang, 2018	(93.7)	(94.46)
WP-[1-4], Args, Joint Learning	**(94.5)**	**(95.33)**

*numbers in parentheses averaged across 10 runs

Results* on Four-way Task

Implicit Relations

Model	Macro-F1	Accuracy	Comparison	Contingency	Expansion	Temporal
WP-[1-4], Args, Implicit Only	**49.2**	**56.11**	42.1	51.1	**64.77**	38.8
WP-[1-4], Args, Joint Learning	**50.2**	**59.13**	41.94	49.81	**69.27**	39.77

*averaged across 10 runs

Results* on Four-way Task

Model	Macro-F1	Accuracy	Macro-F1	Accuracy
	Implicit		Explicit	
Args, Joint Learning	48.1	57.5	94.81	95.63
WP-1, Args, Joint Learning	48.73	57.36	**94.83**	**95.67**
WP-[1-4], Args, Joint Learning	**50.2**	**59.13**	94.50	95.33

*averaged across 10 runs

Results* on Four-way Task

Implicit Relations

Model	Macro-F1	Accuracy	Comparison	Contingency	Expansion	Temporal
Args, Joint Learning	48.1	57.5	**35.5**	52.5	**67.07**	**37.47**
WP-1, Args, Joint Learning	48.73	57.36	**37.33**	52.27	**66.61**	**38.70**
WP-[1-4], Args, Joint Learning	50.2	59.13	**41.94**	49.81	**69.27**	**39.77**

*averaged across 10 runs

Discussion

What types of discourse relations are helped the most by word pairs?

- Comparison (+6.5), Expansion (+2.2), Temporal (+2.3)
- Contingency not helped (-2.7)

Why do word pairs help some classes? Needs more investigation

- Expansion and comparison have words of similar or opposite meaning
- Contingency may benefit more from words indicative of discourse context, e.g. implicit causality verbs (Ronnqvist et al., 2017; Rohde and Horton, 2010)

Qualitative Analysis

1. Removed all non-linearities after convolutional layers
2. Average of 3 runs reduces score from 50.9 to 50.1
3. Argmax of feature maps instead of max pooling
4. Identify examples recovered by joint learning and not by implicit only

Qualitative Analysis

Alliant said it plans to use the microprocessor in future products.

It declined to discuss its plans for upgrading its current product line.

Comparison

Qualitative Analysis

Alliant said it **plans** to use the microprocessor in future products.

It **declined to discuss its plans** for upgrading its current product line.

Comparison

Qualitative Analysis

And it allows Mr. Van de Kamp to get around campaign spending limits

He can spend the legal maximum for his campaign

Expansion

Qualitative Analysis

And it allows Mr. Van de Kamp to get around campaign **spending limits**

He can spend the legal **maximum** for his campaign

Expansion

Model Complexity And Time complexity

- We compare the space and time complexity of our model against two layered Bi-LSTM-CRF model for further comparison.

- We ran each model three times for five epochs to get the wall clock running time

Model	Parameters	Running Time
Ours	1.83M	109.6s
Two layered Bi-LSTM	3.7M	206.17s

Concluding Remarks

- Word pairs are complementary to individual arguments overall and on 3 of 4 first-level classes
- Results on joint learning indicate shared properties of implicit and explicit relations
- Future Work
  - Contextual embeddings
  - External labeled corpora and unlabeled noisy corpora

Questions?

Siddharth Varia: sv2504@columbia.edu

Christopher Hidey: chidey@cs.columbia.edu

Tuhin Chakrabarty: tuhin.chakrabarty@columbia.edu

https://github.com/siddharthvaria/WordPair-CNN

Results on Fifteen-way Task

Model	F1 score			
	Implicit		Explicit	
	PDTB	Blind	PDTB	Blind
Xue et al., (2016)	**40.91**	37.67	**90.22**	**78.56**
Lan et al., (2017)	39.40	**40.12**	-	-
Args, JL	39.68	38.74	89.91	76.98
WP-[1-4], Args, JL	39.39	39.36	89.48	77.00

Related work

- Word Pairs
  - Cross-product of words on either side of the connective (Marcu and Echihabi, 2002; Blair-Goldensohn et al., 2007)
  - Top word pairs are discourse connectives and functional words (Pitler, 2009)
  - Separate TF-IDF word pair features for each connective (Biran and McKeown, 2013)

Pro: large corpus, covers many word pairs

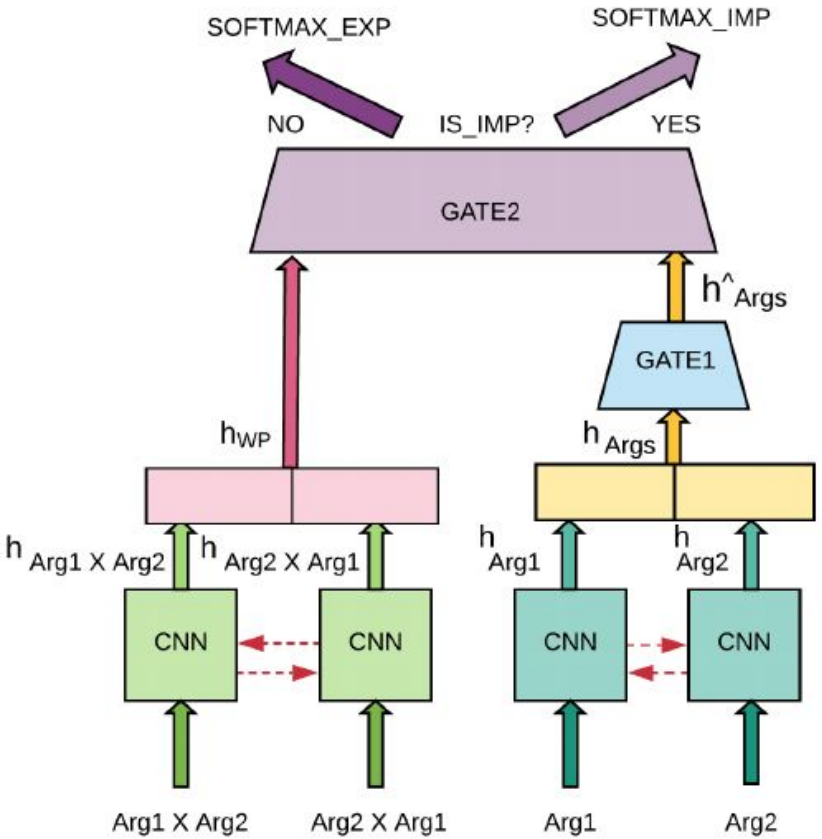
Cons: noisy data, sparsity of word pairs

Neural Models

Pro: easier to transfer knowledge between explicit and implicit

Our Method - 1

- Given the Arguments, Arg1 and Arg2, we learn three types of features from these argument spans:
  - **Word/Word Pairs**
  - **Word/N-Gram Pairs**
  - **N-gram features**
- For first two features, we compute cartesian product of words in Arg1 and Arg2 and feed that as input to convolution layers using filters of sizes 2, 4, 6, 8.
- For N-gram features, we feed the individual arguments Arg1 and Arg2 to second set of convolution layers using filters of sizes 2, 3, 4, 5.



Our Method - 2

- Consider the following sentence:
  - I am late for the meeting because the train was delayed
- Given the phrases “I am late for the meeting” and “the train was delayed”, the cartesian product of words in these two phrases will be as shown in the table on the right
- Each cell in the table is an example of **Word/Word Pair**
- Each row is an example of **Word/N-Gram Pair** where the row word acts as a “**Word**” and the column words act as “**N-gram**”

	the	train	was	delayed
late	late,the	late,train	late,was	late,delayed
for	for,the	for,train	for,was	for,delayed
the	the,the	the,train	the,was	the,delayed
meeting	meeting, the	meeting,t rain	meeting, was	meeting,delaye d

Our Method - 3

- Combination of Argument Representations:
  - As shown in our architecture, we use two identical gates to combine the various features.
- We also perform joint learning of implicit and explicit relations.
- We employ separate softmax classification layers for these two types of relations
- In the nutshell, our architecture is very modular and simple.

$$c_1 = Relu(W_{11} \cdot [\mathbf{h}_{Arg1}; \mathbf{h}_{Arg2}] + b_{11})$$
$$g_{a_1} = \sigma(W_{12} \cdot [\mathbf{h}_{Arg1}; \mathbf{h}_{Arg2}] + b_{12})$$
$$\hat{\mathbf{h}}_{Args} = c_1 \odot g_{a_1}$$
$$c_2 = Relu(W_{21} \cdot [\hat{\mathbf{h}}_{Args}; \mathbf{h}_{WP}] + b_{21})$$
$$g_{a_2} = \sigma(W_{22} \cdot [\hat{\mathbf{h}}_{Args}; \mathbf{h}_{WP}] + b_{22})$$
$$\mathbf{h} = c_2 \odot g_{a_2}$$

Results* on Four-way Task

Implicit Relations

Model	Macro-F1	Accuracy	Comparison	Contingency	Expansion	Temporal
Dai & Huang, 2018	48.82	58.2	37.72	49.39	68.86	**40.7**
WP-[1-4], Args, Implicit Only	49.2	56.11	**42.1**	**51.1**	64.77	38.8
WP-[1-4], Args, Joint Learning	**50.2**	**59.13**	41.94	49.81	**69.27**	39.77

Explicit Relations

Model	Macro-F1	Accuracy
Dai & Huang, 2018	93.7	94.46
WP-[1-4], Args, Joint Learning	94.5	95.33

*averaged across 10 runs

Qualitative Analysis

Alliant said it plans to use the microprocessor in future products
It declined to discuss its plans for upgrading its current product line
Comparison

plans : declined discuss its plans

And it allows Mr. Van de Kamp to get around campaign spending limits
He can spend the legal maximum for his campaign
Expansion

maximum : spending limits