Learning connective-based word representations for implicit discourse relation identification

Chloé Braud and Pascal Denis

CoAStaL
University of Copenhagen

Inria
Lille Nord Europe

EMNLP 2016

Discourse relations

She is in Paris, then she'll move to Copenhagen next week.

It was really cold this week, but it was not as rainy as expected.

Discourse relations

Temporal succession

She is in Paris, then she'll move to Copenhagen next week.

Contrast

It was really cold this week, but it was not as rainy as expected.

Discourse relations

Temporal succession

Arg1 Arg2

[She is in Paris,] then [she'll move to Copenhagen next week.]

Classification task:
x=(Arg1, Arg2) ; y=Temporal
f: X —> Y

Explicit relations

Temporal succession

She is in Paris, **then** she'll move to Copenhagen next week.

Contrast

It was really cold this week, **but** it was not as rainy as expected.

Discourse connectives

Easy task: acc. 94%

Implicit relations

Temporal succession

She is in Paris. ∅ She'll move to Copenhagen next week.

No explicit cue

Hard task: acc. 57%

Half of the relations are implicit

Word pairs features

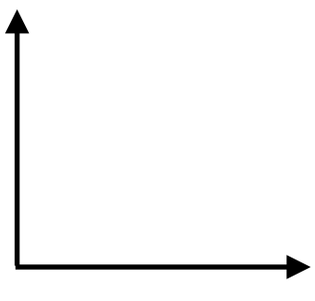
[Marcu and Echihabi 2002]

Arg1: Quarterly revenue **rose** 4.5%.
Arg2: For the year, net income **tumbled** 61%.

Contrast

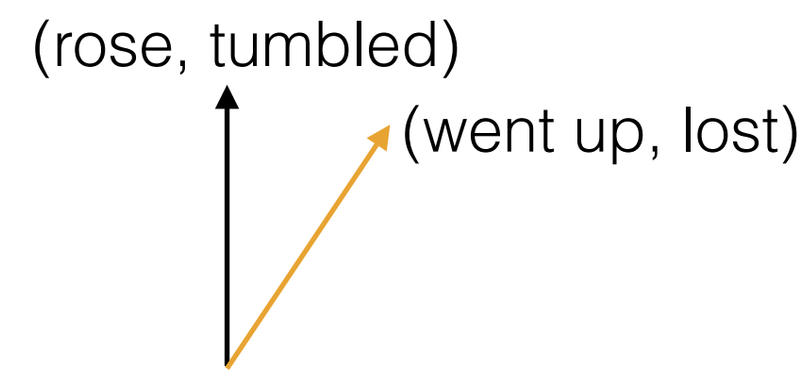
One-hot

(rose, tumbled)



$w \longrightarrow 1_w \in \mathbb{Z}^d,\ d = |V|$

Distributed/distributional representation



$w \longrightarrow v \in \mathbb{R}^d,\ d << |V|$

Previous work

Factors: lexicon, syntax, tense, word knowledge…

Pre-trained **word representations**:

- Not tailored to the (semantic) task

But dense, real-valued

Using **explicit relations** as additional data:

- Adaptation required
- Longer training time

But massive amount of data

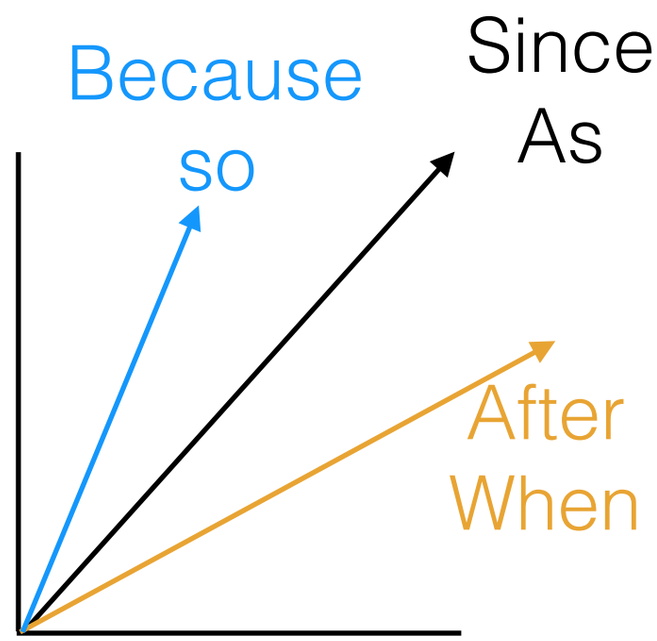
Build a dense representation using the connectives

Discourse-based representation

Assumption: Words occurring in similar *rhetorical* contexts tend to have similar *rhetorical* meanings

Connectives as relevant contexts

- 100 connectives: low dimensional
- Triggering a few relations: keep ambiguity
- Word-based representation: less sparse



Discourse-based representation

She is in Paris, **then** she'll move to Copenhagen **next week**.

More related to Temporal

It was really cold this **week**, **but** it was not as rainy as expected.

Discourse-based representation

Explicit data

w w… **then** w w … week
w w week … **then** w w …
w w week … **but** w w …
w w … **but** w w not …

Frequency counts

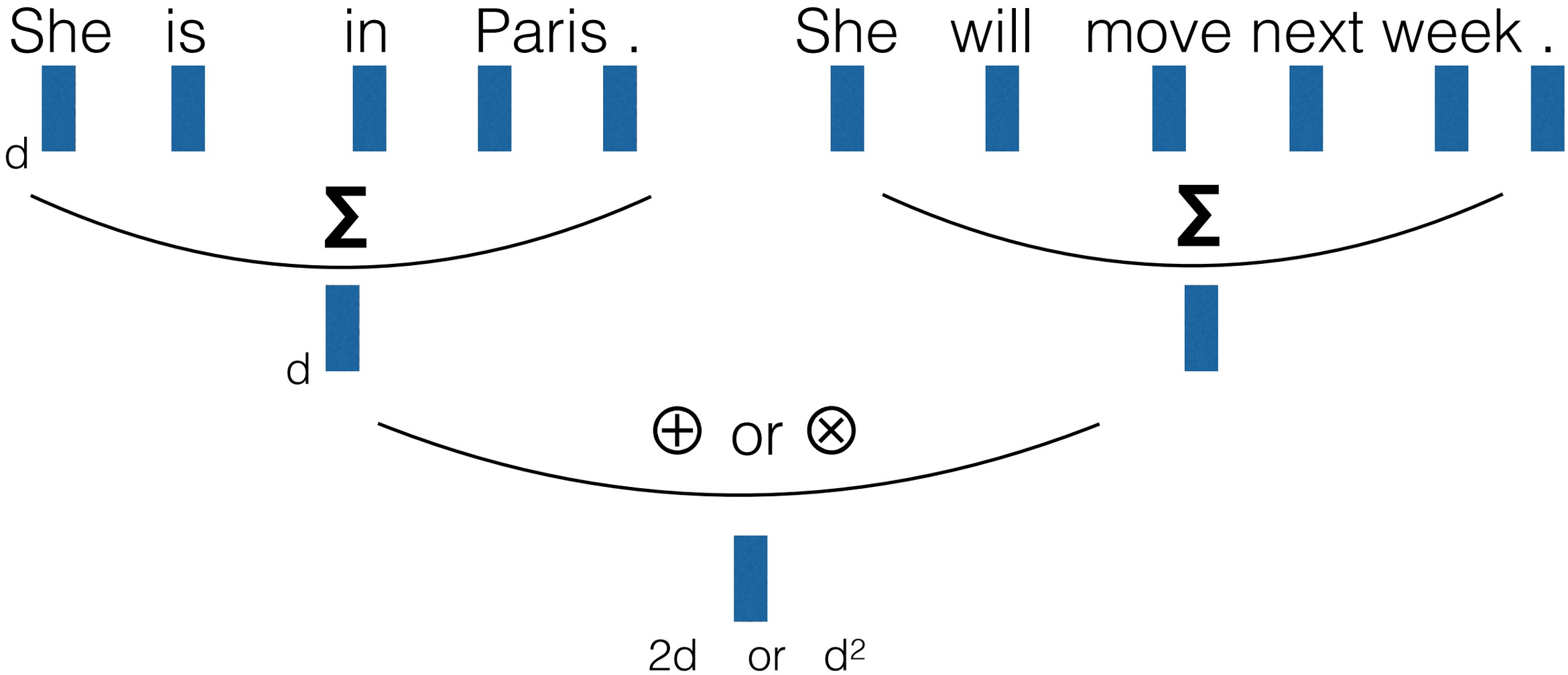
Co-occurence matrix

	but	then
week	0.05	0.3
not	0.4	0.1
…	…	…

Weighting & Normalisation:
TF-IDF or PPMI-IDF
+ PCA

Words are embedded in the connective space

Final representation



Final vector: pairing over the arguments

[Braud and Denis 2015]

Automatic annotation of explicit examples

Identifying the connectives

Use of dispersants was approved **when** a test on the third day showed some positive results, officials said.

As long as your essay is as long as my essay, the teacher will be pleased.

Discourse vs non-discourse

Micro-Acc	Macro-F1
92.9	**91.5**

Automatic annotation of explicit examples

Identifying the arguments

Use of dispersants was approved **when** a test on the third day showed some positive results, officials said.

Such problems will require considerable skill to resolve. **However**, neither Mr. Baum nor Mr. Harper has much international experience.

intra- vs inter-sentential

Micro-Acc	Macro-F1
96.1	**96.0**

Automatic annotation of explicit examples

Identifying the arguments

[Use of dispersants was approved] **when** [a test on the third day showed some positive results,] **officials said.**

[Such problems will require considerable skill to resolve.] **However**, [neither Mr. Baum nor Mr. Harper has much international experience.]

Exact span

3 millions examples automatically extracted from the Bllip 422,199 words in the discourse-based representation

Experiments

Data:

- **Penn Discourse Treebank**
- 4 level-1 relations

PDTB	Train	Test
Temporal	665	68
Contingency	3,281	276
Comparison	1,894	146
Expansion	6,792	556

Model:

- **Multi-class** Logistic Regression
- Class weighting

Experiments

Baselines:

- Word pairs: one-hot ⊕/⊗
- Word pairs: **pre-trained embeddings** ⊕/⊗

  Brown clusters, Collobert and Weston, HLBL, HPCA (Braud and Denis 2015)

Our systems:

- Word pairs: **connective-based embeddings** ⊕/⊗

  TF-IDF or PPMI-IDF + PCA or no PCA

Adding **traditional features**: are they still useful?

Production rules, information on verbs, polarity …

Multi-class Results

Representation	Macro-F1	Micro Acc.
One-hot ⊗	39.0	48.6
One-hot ⊕	40.2	50.2
Braud & Denis 15	41.6	50.1
Braud & Denis 15 +addF	40.8	51.2
Rutherford & Xue 15	40.5	57.1
Bllip TF-IDF ⊗	41.4	51.0
Bllip TF-IDF ⊕	40.1	50.0
Bllip PPMI-IDF ⊗	38.9	48.2
Bllip PPMI-IDF ⊕	**42.2**	52.5
Best Bllip +addF	**42.8**	51.7

—> Best systems: no PCA (ie maximum number of dimensions)

Results per class

	Bllip PPMI-IDF ⊕		Bllip + add feat		Rutherford & Xue	
	Prec.	F1	Prec.	F1	Prec.	F1
Temporal	23.0	**29.9**	23.7	27.9	38.5	14.7
Contingency	49.6	**47.1**	46.7	46.3	49.3	43.9
Comparison	35.9	27.7	35.0	**34.3**	44.9	34.2
Expansion	62.8	64.0	63.7	62.6	61.4	**69.1**

Conclusion

- Representation tailored to the task
- Alleviate the need for external resources

https://bitbucket.org/chloebt/discourse-data

Future work

- Adding new contexts (Alternative lexicalizations, modals, adverbs…)
- More sophisticated weighting schemes
- Directly representing the pairs of words
- Compare to a distributed representation (eg skip-gram)