Comparing Word Representations for Implicit Discourse Relation Classification

Chloé Braud* & Pascal Denis◇

*Alpage, Univ. Paris 7 - INRIA Paris ◇ INRIA Lille

firstname.lastname@inria.fr

Implicit Discourse Relations

- **Implicit relations**: no explicit cues

> S_1 [Quarterly revenue **rose** 4.5%, to \$2.3 billion from \$2.2 billion]
> (whereas/*Comparison*)
> S_2 [For the year, net income **tumbled** 61% to \$86 million, or \$1.55 a share]

- **Complex problem**: lexical, syntactic, temporal, semantic, world knowledge ...,

1. Using a lot of **hand-crafted resources and automatic tools**:
   – **Available but for a few languages** and need pre-processing
2. Using **word-based information in the form of word pairs**:
   $(S_1, S_2) \rightarrow$ < (Quarterly,For), (Quarterly,the), ..., (billion,share) >
   – Easy to build but **one-hot encoding**: **very sparse**

Proposed Strategy

Are unsupervised word representations useful for discourse relation classification?
→ **Dense** representation **available** for virtually any language

Open Questions

1. **Word Representations** What are the most relevant word representations?
   → Compare various word representations: **one-hot**, **cluster-induced** (Rutherford and Xue 2014) or **dense real-valued** (Ji and Eisenstein 2014).
2. **Vector Combination** How to use word representations for a pair of arguments?
   → Compare various ways to build a composite vector: **summation** and **concatenation** (⊕) or **Kroenecker product** (⊗).
3. **Important Words** Are all the words in the segments of equal importance?
   → Compare using **all words** or just **head words**.

Framework

Word Representations

→ Associate a word to a mathematical object, typically a vector in $\{0,1\}^{|\mathcal{V}|}$ or $\mathbb{R}^{|\mathcal{V}|}$, where $\mathcal{V}$ is a base vocabulary

One-hot Word Representations

- Crudest but most common
- Word $w \mapsto \mathbb{1}_w$, d-dimensional indicator vector, $d = |\mathcal{V}|$

Cluster-based One-hot Word Representations

- Learning word representations using hierarchical clustering (Brown et al. 1992)
- Group words in $|\mathcal{C}|$ clusters with $|\mathcal{C}| \ll |\mathcal{V}|$
- Word $w \mapsto \mathbb{1}_w$, k-dimensional indicator vector, $k = |\mathcal{C}|$

Dense Real-Valued Word Representations

- Learning distributed word representations using neural language models (Collobert and Weston 2008, Turian et al. 2010)
- Building distributional word representations using context frequencies and dimensionality reduction, i.e. Hellinger PCA (Lebret and Collobert 2014)
- Represent each word by a vector of p dimensions with $p \ll |\mathcal{V}|$
- Word $w \mapsto \mathbf{v}$, p-dimensional real-valued vector

Vector Combination

→ Generic feature function mapping pairs of segments to a d-dimensional real vector:

$$\Phi : \mathcal{V}^n \times \mathcal{V}^m \rightarrow \mathbb{R}^d, \qquad (S_1, S_2) \mapsto \Phi(S_1, S_2)$$

Representation Based on Head Words

(rose,tumbled) ↦ one vector

- One-hot Representations:
  ▶ $\Phi_{h,1,\oplus}(S_1,S_2) = \mathbb{1}_{\text{rose}} \oplus \mathbb{1}_{\text{tumbled}} \in \{0,1\}^{2|\mathcal{V}_h|}$
  ▶ $\Phi_{h,1,\otimes}(S_1,S_2) = \text{vec}(\mathbb{1}_{\text{rose}} \otimes \mathbb{1}_{\text{tumbled}}) \in \{0,1\}^{|\mathcal{V}_h|^2}$
- Dense Representations:
  ▶ $\Phi_{h,\boldsymbol{M},\oplus}(S_1,S_2) = \boldsymbol{M}^\top \mathbb{1}_{\text{rose}} \oplus \boldsymbol{M}^\top \mathbb{1}_{\text{tumbled}} \in \mathbb{R}^{2p}$
  ▶ $\Phi_{h,\boldsymbol{M},\otimes}(S_1,S_2) = \text{vec}(\boldsymbol{M}^\top \mathbb{1}_{\text{rose}} \otimes \boldsymbol{M}^\top \mathbb{1}_{\text{tumbled}}) \in \mathbb{R}^{p^2}$

$\mathcal{V}_h \subset \mathcal{V}$ the set of head words
$\boldsymbol{M}$ a $n \times p$ real matrix, i^{th} row → p-dimensional embedding of the i^{th} word of $\mathcal{V}_h$

Representation Based on All Words

S_1 [Quarterly revenue rose 4.5%, to \$2.3 billion from \$2.2 billion] ↦ one vector

- Summing over the pairs of words vectors composing the segments

$(S_1 = \{\text{Quaterly}, \ldots, \text{billion}\}, S_2 = \{\text{For}, \ldots, \text{share}\})$ ↦ one vector

- One-hot Representations:
  ▶ $\Phi_{all,1,\oplus}(S_1,S_2) = \sum_i^n \sum_j^m \mathbb{1}_{w_{1_i}} \oplus \mathbb{1}_{w_{2_j}} \in \mathbb{Z}_{\geq 0}^{2|\mathcal{V}|}$
  ▶ $\Phi_{all,1,\otimes}(S_1,S_2) = \sum_i^n \sum_j^m \text{vec}(\mathbb{1}_{w_{1_i}} \otimes \mathbb{1}_{w_{2_j}}) \in \mathbb{Z}_{\geq 0}^{|\mathcal{V}|^2}$
- Dense Representations:
  ▶ $\Phi_{all,\boldsymbol{M},\oplus}(S_1,S_2) = \sum_{i,j}^{n,m} \boldsymbol{M}^\top \mathbb{1}_{w_{1_i}} \oplus \boldsymbol{M}^\top \mathbb{1}_{w_{2_j}} \in \mathbb{R}^{2p}$
  ▶ $\Phi_{all,\boldsymbol{M},\otimes}(S_1,S_2) = \sum_{i,j}^{n,m} \text{vec}(\boldsymbol{M}^\top \mathbb{1}_{w_{1_i}} \otimes \boldsymbol{M}^\top \mathbb{1}_{w_{2_j}}) \in \mathbb{R}^{p^2}$

Experiments

- **Dataset** Penn Discourse Treebank (Prasad et al. 2008), Train: 2-20, Test: 21-22
- **Labels** level 1 relations: *Temporal*, *Contingency*, *Comparison*, *Expansion*
- **Model** MaxEnt + Sample weigthing to deal with class imbalance

F1 score for the best systems using only head words

Repr.	*Temp*	*Cont*	*Comp*	*Expa*
One-hot ⊗	11.96	43.24	17.30	**69.21**
One-hot ⊕	23.01	49.40	29.23	59.08
Brown ⊗	22.91	45.74	25.83	68.76
Brown ⊕	21.84	47.36	27.52	61.38
Embed. ⊗	**23.88**	**51.29**	**30.59**	58.59
Embed. ⊕	22.48	47.48	29.82	57.45

- Heads carry a lot of information
- Using a dense representation is crucial
- Word embeddings are better for heads only

F1 score for the best systems using all words

Repr.	*Temp*	*Cont*	*Comp*	*Expa*
One-hot ⊗	21.14	50.36	34.80	59.43
One-hot ⊕	23.04	51.31	34.06	58.96
Brown ⊗	15.52	**53.85**	30.90	61.87
Brown ⊕	**27.96**	49.48	31.19	**67.42**
Embed. ⊗	22.97	52.76	**34.99**	61.87
Embed. ⊕	25.98	52.50	33.15	60.17

- Need other words: all words give the highest performance
- Brown clusters are better when dealing with all words: could come from the increased number of dimensions to combine or the summation strategy

- **Dense representations are always better**
- **Product is generally better**: keep combination information
- **The best representation is relation dependent**

F1 score for the best systems using all words and extra features

▹ How much improvement can be obtained by **adding** other **standard features**?

- State-of-the-art performance or above when adding extra features
- But improvements are not significant against using only dense representations

Repr.	*Temp*	*Cont*	*Comp*	*Expa*
(Ji and Eisenstein, 2014)	26.91	51.39	35.84	**79.91**
(Rutherford and Xue, 2014)	28.69	54.42	**39.70**	70.23
repr. (Rutherford and Xue, 2014)	24.79	53.39	36.46	50.00
One-hot ⊗ all + add. feats	23.26	54.41	34.34	62.57
Best all + add. feats	**29.30**	**55.76**	36.36	61.76

- **Dense representations already provide most of the semantic and syntactic information relevant to the task**
- **Alleviate the need for traditional external resources**

Perspectives

- Try other combination schemes (Blacoe and Lapata 2012, Le and Mikolov 2014)
- Adapt word representations to the task (Labutov and Lipson 2013, Conrath et al. 2014)