Multi-task Attention-based Neural Networks for Implicit Discourse Relationship Representation and Identification

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, Haifeng Wang

Presented by: Aidan San
Implicit Discourse Relation

“to recognize how two adjacent text spans without explicit discourse marker (i.e., connective, e.g., because or but) between them are logically connected to one another (e.g., cause or contrast)”
Sense Tags

Figure 1: Hierarchy of sense tags.
Implicit Discourse Relation - Motivations

- Discourse Analysis
- Language Generation
- QA
- Machine Translation
- Sentiment Analysis
Summary

- Attention-based neural network conducts discourse relationship representation learning
- Multi-task learning framework leverage knowledge from auxiliary task
Recap - Attention

- Use a vector to scale certain parts of the input so you can “focus” more on that part of the input
Recap - Multi-Task Learning

● Simultaneously train your model on another task to augment your model with additional information

● PS: Nothing crazy in this paper like training with images
Motivation - Attention

- Contrast information can come from different parts of sentence
  - Tenses - Previous vs Now
  - Entities - Their vs Our
  - Whole arguments
- Attention selections most important part of arguments
Motivation - Multi-Task Learning

- Lack of labeled data
- Information from unlabeled data may be helpful
LSTM Neural Network

\[ i_i = \sigma(W_i [x_i, h_{i-1}] + b_i) \]  \hspace{1cm} (3)

\[ f_i = \sigma(W_f [x_i, h_{i-1}] + b_f) \]  \hspace{1cm} (4)

\[ o_i = \sigma(W_o [x_i, h_{i-1}] + b_o) \]  \hspace{1cm} (5)

\[ \tilde{c}_i = \tanh(W_c [x_i, h_{i-1}] + b_c) \]  \hspace{1cm} (6)

\[ c_i = i_i \odot \tilde{c}_i + f_i \odot c_{i-1} \]  \hspace{1cm} (7)

\[ h_i = o_i \odot \tanh(c_i) \]  \hspace{1cm} (8)

Figure 1: LSTM for discourse argument pair representation learning.
Bi-LSTM

Concatenate

Sum-Up Hidden States

\[ h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}] \]

\[ R_{Arg_1} = \sum_{i=1}^{L_1} h_i^1 \quad (9) \]

\[ R_{Arg_2} = \sum_{i=1}^{L_2} h_i^2 \quad (10) \]

Concatenate

\[ R_{pair} = [R_{Arg_1}, R_{Arg_2}] \]
LSTM Neural Network

Figure 1: LSTM for discourse argument pair representation learning.
Attention Neural Network

\[ p_i^1 = \text{Softmax}(R_{Arg2}^T h_i^1) \]  \hspace{1cm} (11)

\[ p_i^2 = \text{Softmax}(R_{Arg1}^T h_i^2) \]  \hspace{1cm} (12)

\[ R'_{Arg1} = \sum_{i=0}^{L_1} h_i^1 p_i^1 \]  \hspace{1cm} (13)

\[ R'_{Arg2} = \sum_{i=0}^{L_2} h_i^2 p_i^2 \]  \hspace{1cm} (14)

\[ R_{pair} = [R'_{Arg1}, R'_{Arg2}] \]

Figure 2: Attention Neural Network for representation learning of arguments.
What is the other task?

- Not really a different task
- Using the explicit data for the same task
Multi-task Attention-based Neural Network

Figure 3: The framework of our proposed multi-task attention-based neural network model.
Knowledge Sharing Methods

1. Equal Share
   \[ \text{Loss} = \text{Loss}_{\text{main}} + \text{Loss}_{\text{aux}} \quad (15) \]

2. Weighted Share
   \[ \text{Loss} = \text{Loss}_{\text{main}} + w \times \text{Loss}_{\text{aux}} \quad (16) \]
   \[ R'_{\text{main}} = R_{\text{main}} \odot \sigma(W_{\text{inter}} R_{\text{aux}} + b_{\text{inter}}) \quad (17) \]

3. Gated Interaction
   \[ R'_{\text{aux}} = R_{\text{aux}} \odot \sigma(W_{\text{inter}} R_{\text{main}} + b_{\text{inter}}) \quad (18) \]
Gated Interaction Cont.

- Acts as a gate to control how much information goes to the end result

\[
R'_{\text{main}} = R_{\text{main}} \odot \sigma(W_{\text{inter}} R_{\text{aux}} + b_{\text{inter}}) \\
R'_{\text{aux}} = R_{\text{aux}} \odot \sigma(W_{\text{inter}} R_{\text{main}} + b_{\text{inter}})
\]

(17)  

(18)

\[
W_{\text{inter}} \in \mathbb{R}^{d_{\text{pair}} \times d_{\text{pair}}} \quad \text{and} \quad b_{\text{inter}} \in \mathbb{R}^{d_{\text{pair}}}
\]

Figure 4: Sigmoid (Gated) interaction shared in multi-task framework (GShare).
Datasets - PDTB 2.0

- Largest Annotated Corpus of discourse relations
- 2,312 Wall Street Journal (WSJ) articles
- Comparison (denoted as Comp.), Contingency (Cont.), Expansion (Exp.) and Temporal (Temp.)
Datasets - CoNLL-2016

- Test - From PDTB
- Blind - From English Wikinews
- Merges labels to remove sparsity
Datasets - BLLIP

- The North American News Text
- Unlabeled data
- Remove Explicit discourse connectives -> Synthetic Implicit Relations
- 100,000 relationships from random sampling
Parameters

- Word2Vec Dimension: 50
- PDTB
  - Hidden State Dimension: 50
  - Multi-task framework hidden layer size: 80
- CoNLL-2016
  - Hidden State Dimension: 100
  - Multi-task framework hidden layer size: 80
Parameters (cont.)

- Dropout: .5 (To penultimate layer)
- Cross-Entropy
- AdaGrad
  - Learning rate: .001
- Minibatch size: 64
Results
|        | Comp. | Cont. | Exp. | Exp+ | Temp  |
|--------|-------|-------|------|------|-------|
| STL    |       |       |      |      |       |
| LSTM   | 33.50 | 52.09 | 67.51| 76.12| 27.88 |
| Bi-LSTM| 33.82 | 52.30 | 67.47| 76.36| 29.01 |
| Attention| **38.15** | **56.07** | **70.53** | **79.80** | **36.72** |
| Eshare |       |       |      |      |       |
| Imp + Exp | 35.07 | 54.62 | 69.97| 79.15| 34.57 |
| Imp + BLLIP | 37.67 | 56.82 | 70.81| 80.43| 35.48 |
| Wshare |       |       |      |      |       |
| Imp + Exp | 37.51 (w=0.1) | 55.83 (w=0.2) | 70.37 (w=0.3) | 80.22 (w=0.2) | 35.71 (w=0.3) |
| Imp + BLLIP | 39.13 (w=0.2) | 57.78 (w=0.2) | 71.88 (w=0.1) | 80.84 (w=0.3) | 37.76 (w=0.3) |
| Gshare |       |       |      |      |       |
| Imp + Exp | 38.91 | 56.91 | 71.41| 80.02| 36.92 |
| Imp + BLLIP | **40.73** | **58.96** | **72.47** | **81.36** | **38.50** |

Table 2: Performance of multiple binary classification on the top level classes in PDTB corpus in terms of $F_1$ (%).
| STL       | PDTB (Four way)                                      | CoNLL-Test (Acc) | CoNLL-Blind (Acc) |
|-----------|------------------------------------------------------|------------------|-------------------|
| LSTM      | $F_1$: 36.16; Acc: 56.12                            | 34.45            | 35.07             |
| Bi-LSTM   | $F_1$: 36.54; Acc: 54.30                            | 34.85            | 35.83             |
| Attention | $F_1$: 45.57; Acc: 57.55                            | **37.41**        | **38.36**         |
| Eshare    | $F_1$: 44.17; Acc: 55.65                            | 35.56            | 37.06             |
| Imp + Exp | $F_1$: 44.57; Acc: 55.85                            | 36.66            | 38.28             |
| Imp + BLLIP |                                                  |                  |                   |
| Wshare    | $F_1$: 45.03; Acc: 56.21 ($w=0.3$)                    | 36.24 ($w=0.2$)  | 37.34 ($w=0.3$)   |
| Imp + Exp | $F_1$: 45.80; Acc: **58.95** ($w=0.2$)                | 38.13 ($w=0.1$)  | 39.14 ($w=0.4$)   |
| Imp + BLLIP |                                                  |                  |                   |
| Gshare    | $F_1$: 45.70; Acc: 57.17                            | 37.84            | 38.10             |
| Imp + Exp | $F_1$: **47.80**; Acc: 57.39                         | **39.40**        | **40.12**         |
| Imp + BLLIP |                                                  |                  |                   |

Table 3: Performance of multi-class classification on PDTB and CoNLL-2016 in terms of accuracy (Acc) (%) and macro-averaged $F_1$ (%).
|                         | Binary Classification ($F_1$) | Multi-class Classification (Acc) |
|-------------------------|-------------------------------|----------------------------------|
|                         | Comp. | Cont. | Exp. | Exp+ | Temp | PDTB (Four way) | CoNLL-Test(Acc) | CoNLL-Blind(Acc) |
| (Chen et al., 2016)     | 40.17 | 54.76 | -    | 80.62| 31.32| -                | -                | -                |
| (Qin et al., 2016b)     | 41.55 | 57.32 | 71.50| 80.96| 35.43| -                | -                | -                |
| (Liu and Li, 2016)      | 39.86 | 54.48 | 70.43| 80.86| 38.84| -                | -                | -                |
| (Wu et al., 2016)       | -     | -     | -    | -    | -    | -                | -                | -                |
| (Qin et al., 2016a)     | 38.67 | 54.91 | -    | 80.66| 32.76| -                | -                | -                |
| (Liu et al., 2016b)     | 37.91 | 55.88 | 69.97| -    | 37.17| -                | -                | -                |
| (Lan et al., 2013)      | 31.53 | 47.52 | 70.01| -    | 29.51| -                | -                | -                |
| (Wang and Lan, 2016)    | -     | -     | -    | -    | -    | -                | -                | -                |
| (Rutherford and Xue, 2016)| -   | -     | -    | -    | -    | -                | -                | -                |
| Our model               | 40.73 | 58.96 | 72.47| 81.36| 38.50| $F_1$: 47.80; Acc: 57.39 | 39.40            | 40.12            |

Table 4: Comparison with the state-of-the-art systems reported on PDTB and CoNLL-2016, where - means N.A.
Effect of Weight Parameter

Low value of $W$ reduces weight of auxiliary task and makes model pay more attention to main task.
Conclusion

- Multi-task attention-based neural network
- Implicit discourse relationship
- Discourse arguments and interactions between annotated and unannotated data
- Outperforms state-of-the-art