A Systematic Study of Neural Discourse Models for Implicit Discourse Relation

Attapol T. Rutherford. Vera Demberg  Nianwen Xue

Presenter: Dhruv Agarwal
INTRODUCTION

• Inferring implicit discourse relations is a difficult subtask in discourse parsing.

• Typical approaches have used hand crafted features from the two arguments and suffer from data sparsity problems.

• Neural network approaches need to be applied on small datasets and possess no common experimental settings for evaluation.

• This paper conducts several experiments to compare various neural architectures in literature and publishes their results.
DISCOURSE

• High level organization of text can be characterized as discourse relations between adjacent pairs of texts.

• There are two types of discourse relations:
  • Explicit Discourse Relations
  • Implicit Discourse Relations
According to Lawrence Eckenfelder, a securities industry analyst at Prudential-Bache Securities Inc., "Kemper is the first firm to make a major statement with program trading." He added that "having just one firm do this isn't going to mean a hill of beans. But if this prompts others to consider the same thing, then it may become much more important."

The discourse connective is 'but', and the sense is Comparison.Concession.
According to Lawrence Eckenfelder, a securities industry analyst at Prudential-Bache Securities Inc., "Kemper is the first firm to make a major statement with program trading." He added that "having just one firm do this isn't going to mean a hill of beans. But if this prompts others to consider the same thing, then it may become much more important."

The omitted discourse connective is 'however'. and the sense is Comparison.Contrast.
CHALLENGES

• Predicting internal discourse relations is fundamentally a semantic task and relevant semantics might be difficult to recover from surface level features.

  *Bob gave Tina the burger.*
  *She was hungry.*

• Purely vector based representations of arguments might not be sufficient to capture discourse relations.

  *Bob gave Tina the burger.*
  *He was hungry.*
MODEL ARCHITECTURES

• In order to find the best distributed representation and network architecture, they explore by probing the different points on the spectrum of structurality from structureless bag-of-words models to sequential and tree-structured models.

• Bag of words Feed Forward Model

• Sequential LSTM

• Tree LSTM
THREE KINDS OF POOLING ARE CONSIDERED: MAX, MEAN AND SUMMATION AS FOLLOWS,

\[
f_{\text{max}}(w_{1:N}, i) = \max_{j=1}^{N} w_{j,i}
\]

\[
f_{\text{sum}}(w_{1:N}, i) = \sum_{j=1}^{N} w_{j,i}
\]

\[
f_{\text{mean}}(w_{1:N}, i) = \frac{1}{N} \sum_{j=1}^{N} w_{j,i}
\]
• The difference between standard LSTM and Tree LSTM is that gating vectors and memory cell updates are based on hidden states of many child nodes.
• The hidden state vector correspond to a constituent in the tree.
IMPLEMENTATION DETAILS

- Penn Discourse Tree Bank is used because of its theoretical simplicity and large size.

- The PDTB provides three levels of discourse relations, each level providing finer semantic distinctions.

- The task is carried out on the second level with 11 classes.

- Cross Entropy Loss function, Adagrad Optimizer and no regularization/dropout.

- The model performance is also evaluated on CONLL shared task 2015 and CDTB.
### RESULTS

- **FEEDFORWARD MODEL IS THE BEST OVERALL AMONG ALL THE NEURAL ARCHITECTURES THEY EXPLORE.**
- **IT OUTPERFORMS LSTM BASED, CNN BASED AND THE BEST MANUAL SURFACE FEATURE BASED MODELS IN SEVERAL SETTINGS.**

| Model                                      | Accuracy |
|--------------------------------------------|----------|
| PDTB Second-level senses                   |          |
| Most frequent tag baseline                 | 25.71    |
| Our best tree LSTM                         | 34.07    |
| Ji & Eisenstein, (2015)                    | 36.98    |
| Our best sequential LSTM variant           | 38.38    |
| Our best feedforward variant               | 39.56    |
| Lin et al., (2009)                         | 40.20    |

| Systems                          | Arg vector       | Features? | Blind set | WSJ Test | WSJ Dev |
|----------------------------------|------------------|-----------|-----------|----------|---------|
| Ours                             | Summing vectors  | No        | 0.3767    | 0.3613   | 0.4032  |
| Akanksha & Eisenstein (2016)     | 2-layer Bi-LSTM  | Yes       | 0.3675    | 0.3495   | 0.4072  |
| Qin et al. (2016)                | Convolutional net| No        | 0.3538    | 0.3820   | 0.4632  |
| Mihaylov & Frank (2016)          | Averaging vectors| Yes       | 0.3451    | 0.3919   | 0.4032  |
| Schenk et al. (2016)             | Avg + Product    | No        | 0.3185    | 0.3761   | 0.4542  |
| Wang & Lan (2016)                | Convolutional net| No        | 0.3418    | **0.4091**| **0.4642**|
| Wang & Lan (2015)                | N/A              | Yes       | 0.3629    | 0.3445   | 0.4272  |
- For Baseline comparison Max Entropy models are used which are loaded with feature sets such as dependency rule pairs, production rule pairs and Brown Cluster pairs.
Sequential LSTMs outperform feedforward model when word vectors are not high dimensional and not trained on a larger corpus.

Summation pooling is effective for both LSTM and feedforward models, since word vectors are known to have additive properties.
DISCUSSION

• Sequential and Tree LSTM might work better if they had a larger amount of annotated data.

• Benefits of Tree LSTM cannot be realized if the model discards syntactic categories in intermediate nodes.

• Linear interaction allows combination of high dimensional vectors without exponential growth of parameters.
CONCLUSION

- Manually crafted surface features are not important for this task and it holds true for different languages.

- Expressive power of distributed representations can overcome data sparsity issues of traditional approaches.

- Simple feed-forward architecture can outperform more sophisticated architectures such as sequential and tree-based LSTM networks, given the small amount of data.

- The paper compiles the results of all the previous systems and provides a common experimental setting for future research.
THANK YOU