Robust Non-Explicit Neural Discourse Parser in English and Chinese

Attapol T. Rutherford*
Yelp
San Francisco, CA, USA
teruth@yelp.com

Nianwen Xue
Brandeis University
Waltham, MA, USA
xuen@brandeis.edu

Abstract

Neural discourse models proposed so far are very sophisticated and tuned specifically to certain label sets. These are effective, but unwieldy to deploy or repurpose for different label sets or languages. Here, we propose a robust neural classifier for non-explicit discourse relations for both English and Chinese in CoNLL 2016 Shared Task datasets. Our model only requires word vectors and simple feed-forward training procedure, which we have previously shown to work better than some of the more sophisticated neural architecture such as long-short term memory model. Our Chinese model outperforms feature-based model and performs competitively against other teams. Our model obtains the state-of-the-art results on the English blind test set, which is used as the main criteria in this competition.

1 Introduction

In the context of CoNLL 2016 Shared Task, we participate partially in the English and Chinese supplementary evaluation, which is discourse relation sense classification (Xue et al., 2016). We focus on identifying the sense of non-explicit discourse relations in both English and Chinese. Previous studies including the results from CoNLL 2015 Shared Task have shown that classifying the senses of implicit discourse relations is the most difficult part of the task of discourse parsing (Xue et al., 2015). Therefore, we focus exclusively on this particular challenging subtask.

We want our system to be robust such that the system can be easily trained to handle different label sets and different languages. Neural network is attractive in this regard as we do not need handcrafted linguistic resources, which are not readily available in all languages. The past neural network models for this task focus on top-level senses (Ji et al., 2016) or require parses (Ji and Eisenstein, 2015), redundant surface features (Rutherford and Xue, 2014), or extensive semantic lexicon (Pitler et al., 2009). The results from these systems are not likely to extend to languages that do not have as much linguistic resources as English. Therefore, we come up with a neural network model that requires no parses and specific model tuning. The only extra ingredient is word vectors, which are easily obtained through large amount of unannotated data.

Our past studies have indicated that feedforward neural networks outperform more complicated models such as long-short term memory models and perform comparably with systems with traditional surface features in this task (Rutherford et al., 2016). But we want to test our results further. We wonder whether our best feedforward architecture can be adopted to deal with a totally different language and a different label set put forth specifically for this shared task. We also want to know whether our model is robust against the slightly out-of-domain blind datasets.

The performance numbers from the experiments alone hardly provide us with insight into implicit discourse relations. We compare and contrast the two approaches in more detail to learn what we gain and lose by using each approach. The fundamental difference between our approach and the baseline is that our approach does not use surface features or semantic lexicons. We want to know the advantage one gains from shifting the paradigm from discrete surface features to continuous features. Are the errors made by two types of systems complementary?

*Work performed while being a student at Brandeis

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 55–59,
Berlin, Germany, August 7-12, 2016.

Our system is ranked the first on the English dataset and the third on the Chinese dataset. The accuracy on the English blind test set is 0.3767, and the accuracy on the Chinese blind test set is 0.6338. The performance on the test sets even exceeds the one on the development sets, which suggest the robustness of our model.

2 Model description

The Arg1 vector a^1 and Arg2 vector a^2 are computed by applying element-wise pooling function f on all of the N_1 word vectors in Arg1 $w^1_{1:N_1}$ and all of the N_2 word vectors in Arg2 $w^2_{1:N_2}$ respectively:

$$a^1_i = \sum_{j=1}^{N} w^1_{j,i}$$

$$a^2_i = \sum_{j=1}^{N} w^2_{j,i}$$

Inter-argument interaction is modeled directly by the hidden layers that take argument vectors as features. Discourse relations cannot be determined based on the two arguments individually. Instead, the sense of the relation can only be determined when the arguments in a discourse relation are analyzed jointly. The first hidden layer h_1 is the non-linear transformation of the weighted linear combination of the argument vectors:

$$h_1 = \tanh(W_1 \cdot a^1 + W_2 \cdot a^2 + b_{h_1})$$

where W_1 and W_2 are $d \times k$ weight matrices and b_{h_1} is a d-dimensional bias vector. Further hidden layers h_t and the output layer o follow the standard feedforward neural network model.

$$h_t = \tanh(W_{h_t} \cdot h_{t-1} + b_{h_t})$$

$$o = \text{softmax}(W_o \cdot h_T + b_o)$$

where W_{h_t} is a $d \times d$ weight matrix, b_{h_t} is a d-dimensional bias vector, and T is the number of hidden layers in the network.

We think that this model architecture should be effective because we have run extensive studies and experiments on many configuration and architectures (Rutherford et al., 2016). We have experimented and tuned most components: pooling functions for the argument vectors, the type of word vectors, and the model architectures themselves. We found the model variant with two hidden layers and 300 hidden units to work well across many settings. The model has the total of around 270k parameters.

3 Experiments

Word vectors English word vectors are taken from 300-dimensional Skip-gram word vectors trained on Google News data, provided by the shared task organizers (Mikolov et al., 2013; Xue et al., 2015). We trained our own 250-dimensional Chinese word vectors on Gigaword corpus, which is the same corpus used by the 300-dimensional Chinese word vectors provided by the shared task organizers (Graff and Chen, 2005). We found the 250-dimensional version to work better despite fewer parameters.

Training Weight initialization is uniform random, following the formula recommended by Bengio (2012). Word vectors are fixed during training. The cost function is the standard cross-entropy loss function, and we use Adagrad as the optimization algorithm of choice. We monitor the accuracy on the development set to determine convergence.

Implementation All of the models are implemented in Theano (Bergstra et al., 2010; Bastien et al., 2012). The gradient computation is done with symbolic differentiation, a functionality provided by Theano. The models are trained on CPUs on Intel Xeon X5690 3.47GHz, using only a single core per model. The models converge in minutes. The implementation, the training script, and the trained model are already made available [1] .

Baseline The winning system from last year's task serves as a strong baseline for English. We choose this system because it represents one of the strongest systems that utilizes exclusively surface features and extensive semantic lexicon (Wang and Lan, 2015). This approach uses a MaxEnt model loaded with millions of features.

We use Brown cluster pair features as the baseline for Chinese as there is no previous system for Chinese. We use 3,200 clusters to create features and perform feature selection on the development set based on the information gain criteria (Rutherford and Xue, 2014). We end up with 10,000 features total.

4 Results and Discussion

The English results are summarized in Table 1. The English baseline we use is from the winning system from last year's task (Wang and Lan, 2015). Our system is more accurate than the baseline on the two test sets but not on the develop-

[1] https://github.com/attapol/nn_discourse_parser

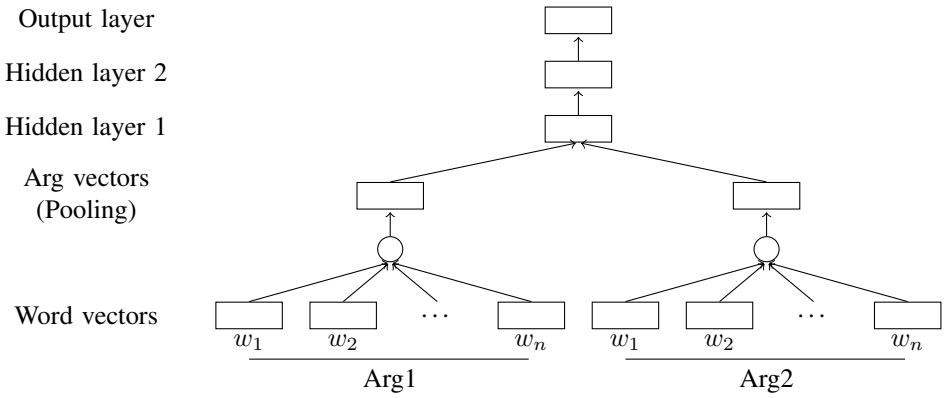

Figure 1: Model architecture

Sense	Development set Baseline	Development set Ours	Test set Baseline	Test set Ours	Blind test set Baseline	Blind test set Ours
Comparison.Concession	0	0	0	0	0	0
Comparison.Contrast	0.098	0.1296	0.1733	0.1067	0	0
Contingency.Cause.Reason	0.4398	0.3514	0.3621	0.4	0.2878	0.3103
Contingency.Cause.Result	0.2597	0.1951	0.1549	0.1722	0.2254	0.1818
EntRel	0.6247	0.5613	0.5265	0.4892	0.5471	0.5516
Expansion.Alternative.Chosen alternative	0	0	0	0	0	0
Expansion.Conjunction	**0.4591**	0.3874	**0.3068**	0.2468	**0.3154**	0.2644
Expansion.Instantiation	0.2105	**0.4051**	0.3261	**0.4962**	0.1633	**0.25**
Expansion.Restatement	0.3482	0.3454	0.2923	0.3483	0.3232	0.2991
Temporal.Asynchronous.Precedence	0	0.0714	0	0	0	0.125
Temporal.Asynchronous.Succession	0	0	0	0	0	0
Temporal.Synchrony	0	0	0	0	0	0
Accuracy	0.4331	0.4032	0.3455	0.3613	0.3629	0.3767
Most-frequent-tag Acc.	0.2320		0.2844		0.2136	

Table 1: F_1 scores for English non-explicit discourse relation. The bold-faced numbers highlight the senses where the classification of our model and the baseline model might be complementary.

ment set. Both systems only learn the top six or seven senses because the other senses constitute only around 5% of the training set, which might not be enough when compared to the complexity of the task.

Our system outperforms the most frequent tag baseline and Brown cluster pair baseline by 7% and by 3% (absolute) respectively in the CDTB datasets (Table 2). Our system only learns to distinguish between EntRel, Conjunction, and Expansion, which are the top three most frequent senses in the training set. The fourth most frequent class, Causation, constitute only around 200 instances in the training set, which is too small for machine learning approaches.

Generally, we would expect the performance on the in-domain test set to be worse than the performance on the in-domain development set. However, we do not observe this trend in the Chinese evaluation. This suggests that our model shows some robustness. Similarly, we would expect the performance on the slightly-out-of-domain test set to be worse than the performance on the in-domain test set. This is also not the case for the English data, which suggests robustness of the model.

What is the trade-off in terms of the performance? The results suggests that the two approaches are partially complementary at least for English. For example, our system does significantly better on Expansion.Instantiation, but the surface feature system does significantly better on Expansion.Conjunction (Table 1). This suggests that surface feature approach still holds some advantage over the neural network approach that we propose here. In the next section, we compare the errors each of the systems more quantitatively.

5 Error Analysis

Comparing confusion matrices from the two approaches help us understand further what neural

	Development set		Test set		Blind test set	
Sense	Baseline	Ours	Baseline	Ours	Baseline	Ours
Alternative	0	0	0	0	0	0
Causation	0	0	0	0	0	0
Conditional	0	0	0	0	0	0
Conjunction	0.7830	0.7928	0.7911	0.8055	0.7875	0.7655
Contrast	0	0	0	0	0	0
EntRel	0.4176	0.4615	0.5175	0.5426	0.0233	0.0395
Expansion	0.4615	0.4167	0.2333	0.4333	0.2574	0.5104
Purpose	0	0	0	0	0	0
Temporal	0	0	0	0	0	0
Accuracy	0.6634	0.683	0.6657	0.7047	0.6437	0.6338
Most-frequent-tag Acc.	0.6176		0.6351		0.7914	

Table 2: F_1 scores for Chinese non-explicit discourse relation.

	The true sense is ...				
, confused as ... by ...	Instantiation	Contrast	Result	Precedence	Synchrony
Conjunction		+		#+	#+
Restatement	+				
Result		#			#
Reason			+		

Table 3: Confusion pairs made by our neural network (#) and the baseline surface features (+) in English.

networks have achieved. We approximate Bayes Factors with uniform prior for each sense pair (c_i, c_j) for gold standard g and system p:

$$\frac{P(p = c_i, g = c_j)}{P(p = c_i)P(g = c_j)}$$

We tabulate all significant confusion pairs (i.e. Bayes Factor greater than a cut-off) made by each of the systems (Table 3). This is done on the development set only.

The distribution of the confusion pairs suggest that neural network and surface feature systems complement each other in some way. We see that the two systems only share two confusion pairs in common.

Temporal.Asynchronous senses are confused with Conjunction by both systems. Temporal senses are difficult to classify in implicit discourse relations since the annotation itself can be quite ambiguous. Expansion.Instantiation relations are misclassified as Expansion.Restatement by surface feature systems. Neural network system performs better on Expansion.Instantiation than surface feature systems probably because neural network system can tease apart Expansion.Instantiation and Expansion.Restatement.

6 Conclusions

We present a robust neural network model, which is easy to deploy, retrain, and adapt to other languages and label sets. The model only needs word vectors trained on large corpora, which are available in most major languages. Our approach performs competitively if not better than traditional systems with surface features and MaxEnt model despite having one or two orders of magnitude fewer parameters. Our results suggest that simple feedforward architecture can be more powerful than more sophisticated neural architectures undertaken by other systems in this shared task.

References

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. 2012. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

Yoshua Bengio. 2012. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*, pages 437–478. Springer.

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June. Oral Presentation.

David Graff and Ke Chen. 2005. Chinese gigaword. *LDC Catalog No.: LDC2003T09, ISBN*, 1:58563–58230.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 683–691. Association for Computational Linguistics.

Attapol T. Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, Gothenburg, Sweden, April.

A. T. Rutherford, V. Demberg, and N. Xue. 2016. Neural Network Models for Implicit Discourse Relation Classification in English and Chinese without Surface Features. *ArXiv e-prints*, June.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24, Beijing, China, July. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China, July. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Bonnie Webber, Attapol Rutherford, Chuan Wang, and Hongmin Wang. 2016. The conll-2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.