Discourse Sense Classification from Scratch using Focused RNNs

Gregor Weiss, Marko Bajec
University of Ljubljana
Faculty of Computer and Information Science
Večna pot 113, Ljubljana, Slovenia
gregor.weiss@student.uni-lj.si
marko.bajec@fri.uni-lj.si

Abstract

The subtask of CoNLL 2016 Shared Task focuses on sense classification of multilingual shallow discourse relations. Existing systems rely heavily on external resources, hand-engineered features, patterns, and complex pipelines fine-tuned for the English language. In this paper we describe a different approach and system inspired by end-to-end training of deep neural networks. Its input consists of only sequences of tokens, which are processed by our novel focused RNNs layer, and followed by a dense neural network for classification. Neural networks implicitly learn latent features useful for discourse relation sense classification, make the approach almost language-agnostic and independent of prior linguistic knowledge. In the closed-track sense classification task our system achieved overall 0.5246 F_1-measure on English blind dataset and achieved the new state-of-the-art of 0.7292 F_1-measure on Chinese blind dataset.

1 Introduction

Shallow discourse parsing is a challenging natural language processing task and sense classification is its most difficult subtask (Lin et al., 2014; Xue et al., 2015). Given text spans for argument 1 and 2, connective, and punctuation, the goal is to predict the sense of the discourse relation that holds between them. These text spans can appear in various orders, are not necessarily continuous, can spread across multiple sentences, and sometimes connectives and punctuation are not even present. The CoNLL 2016 Shared Task (Xue et al., 2016) focuses on multilingual shallow discourse parsing based on the English Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) and Chinese Discourse TreeBank (CDTB) (Zhou and Xue, 2012). Evaluation is performed on separate test and blind datasets on the remote TIRA evaluation system (Potthast et al., 2014).

Existing systems for discourse parsing rely heavily on existing resources, hand-engineered features, patterns, and complex pipelines fine-tuned for the English language (Xue et al., 2015; Wang and Lan, 2015; Stepanov et al., 2015). Such features include word lists, part-of-speech tags, chunking tags, syntactic features extracted from constituent parse trees, path features built around connectives or specific words, production rules, dependency rules, Brown cluster pairs, features that disambiguate problematic connectives, and similar. Similar to our system, these pipelines separately process explicit and non-explicit discourse relation types.

In this paper we describe a different approach and system inspired by end-to-end training of deep neural networks. Instead of engineering features and incorporating linguistic knowledge into them, its input consists of only sequences of tokens. They are processed by a neural network model that utilizes our novel focused recurrent neural networks (RNNs). It automatically learns latent features and how to allocate focus for our task. This way the system is independent of any prior knowledge, existing parsers, or external resources, what makes it almost language-agnostic. By only changing a few hyper-parameters, we successfully applied the same system to the English and Chinese datasets and achieved new state-of-the-art results on the Chinese blind dataset. Our system[1] was developed in Python using the Keras library (Chollet, 2015) that enables it to run on either CPU or GPU.

[1] http://github.com/gw0/conll16st-v34-focused-rnns/

The system architecture is described in Section 2, followed by details of layers in our neural network and their training. Section 3 presents official evaluation results on English and Chinese datasets. Section 4 draws conclusions and directions for future work.

2 System Overview

Our system for discourse sense classification of the CoNLL 2016 Shared Task consists of two similar neural network models build from three types of layers (see Figure 1). In the spirit of end-to-end training its input consists of only tokenized text spans that are mapped to vocabulary ids, which are processed by our neural network to classify each discourse relation into a sense category.

Important steps of our system are:

- **Two models** for separately handling present and absent connectives in discourse relations.
- **Input** consists of four sequences of tokens mapped to vocabulary ids (for argument 1 and 2, connectives, and punctuations).
- **Word Embeddings** layer maps each token into a low-dimensional vector space using a lookup table.
- **Focused RNNs** layer focuses multiple RNNs onto different aspects of these sequences.
- **Classification** is performed with a dense neural network and logistic regression on top.

We used the same system on the English and Chinese datasets and each one uses two separate neural network models with only a few differences in its 18 parameters. Because of these differences, individual models are trained and applied completely separately, although parts could be shared. Total number of trainable weights for both neural network models is 1355661/1185006 for English and 369972/1276761 for Chinese.

2.1 Two models

According to suggestions from related work we separately handle discourse relations with and without given connectives. For each case we train a separate neural network model with the same architecture, but different hyper-parameters. Throughout the paper we present those differences in parameters with a/b, where a presents a value used for Explicit and AltLex relation types (where connectives are present) and b for Implicit and EntRel relation types (where connectives are absent).

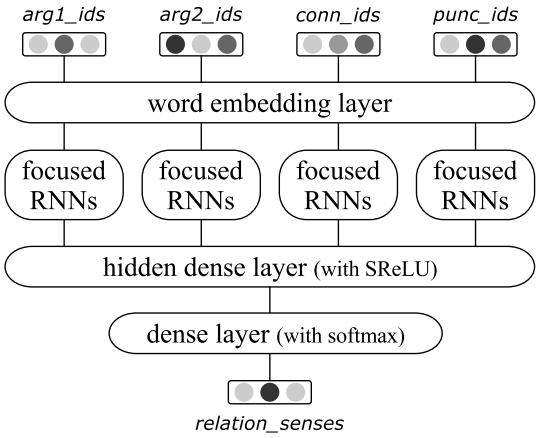

Figure 1: Our neural network model for end-to-end training of sense classification. Two such models are separately trained for each language.

2.2 Input

Initially a vocabulary of all words or tokens in the training dataset is prepared mapping each one to a unique token id. Four text spans representing individual shallow discourse relations are tokenized and mapped into four sequences of vocabulary ids. Depending on the language these input sequences are cropped to different maximal lengths, see Table 1. Out-of-vocabulary words that are not present during training are mapped to a special id.

Relation part	English	Chinese
Argument 1	100	500
Argument 2	100	500
Connective	10	10
Punctuation	2	2

Table 1: Maximal lengths of input sequences in our system for English and Chinese datasets.

2.3 Word embeddings

A shared word embedding layer turns previous sequences of positive integers (token ids) into dense vectors of fixed size using a lookup table. These vector representations are automatically learned with the rest of the model using backpropagation. All four input sequences are mapped into the same low-dimensional vector space with 30/20 dimensions for English and 20/70 for Chinese. For regu-

larization purposes we randomly drop embeddings during training with probability 0.1.

Although the closed-track allowed the use of pre-trained skip-gram neural word embeddings (Mikolov et al., 2013), we decided to learn them from scratch for each model separately.

2.4 Focused RNNs

These embeddings are processed by our novel focused RNNs layer. Any recurrent neural network (RNN) can be used as its building block, but we decided to use the GRU layer (Chung et al., 2014). First a special focus RNN with 4/6 dimensions for English and 4/5 for Chinese is used to assign multidimensional focus weights to the input sequence. For each focus dimension a separate RNN is applied to the input sequence multiplied with corresponding focus weights. This way different RNNs can focus on different aspects of input sequences– in our case on different words and senses. Final outputs of these RNNs are concatenated and used in the classification layers. Our system uses separate RNNs with 10/50 dimensions for English and 20/30 for Chinese. For regularization purposes we randomly drop 0.33 input gates of focus and separate RNNs, 0.66 recurrent connections of the focus RNN, and 0.33 of separate RNNs.

Note that our focused RNNs layer differs a lot from other attention mechanisms found in literature. They are designed to only work with question-answering systems, use a weighted combination of all input states, and can focus on only one aspect of the input sequence.

2.5 Classification

Classification into discourse sense categories is performed using a dense neural network. Merged outputs of all focused RNNs are first processed by a dense layer with 90/40 dimensions for English and 100/90 for Chinese, followed by the SReLU activation function (Jin et al., 2015). The S-shaped rectified linear activation unit (SReLU) consists of piecewise linear functions and can learn both convex and non-convex functions. Finally logistic regression, i.e. a dense layer followed by the softmax activation function, is applied to get classification probabilities. For regularization purposes we randomly drop connections before the second dense layers with probability 0.5.

2.6 Training

Loss function suitable for our classification task is the categorical cross-entropy. Training is achieved with backpropagation and any gradient descent optimization, such as Adam optimizer. To parallelize and speed up the learning process we train in batches of 64 training samples. During training we monitor the loss function on the validation dataset and stop if it does not increase in the last 20 epochs. For regularization purposes we also introduce 32 random noise samples for each discourse relation during training. Weights used by the resulting system are those with the best encountered validation loss.

3 Evaluation

Datasets used by the CoNLL 2016 Shared Task consist of PDTB for English, CDTB for Chinese, and two unknown blind test datasets from Wikinews. For each language there is a train dataset for training models, validation dataset for monitoring the learning process, and test and blind test datasets for evaluating its performance.

Metric used for this subtask of CoNLL 2016 Shared Task is the F_1-measure. It is computed based on the number of predicted discourse relation senses that match a gold standard relation.

3.1 Results for English

The training dataset from PDTB for English consists of 1756 documents with 15246 discourse relations that can be categorized into 15 different discourse relation senses.

Overall our system performs pretty well on all English datasets (see Table 2) despite not using any external resources or hand-engineered features. As expected it performs best on the validation dataset, achieves slightly lower scores (0.5845) on the test dataset, and performs the worst on the blind dataset (0.5246) that contains a different writing style than PDTB. For only explicit relations our system performs much better, close to inter-annotator agreement (91%) on development and test datasets, but without using any word lists or patterns like other systems. On the other hand non-explicit relations seem to be a much harder problem and the relatively small size of the training dataset does not contain enough information.

Detailed per-sense analysis on all discourse relations is shown in Table 3. We see

Type	Dev	Test	Blind
Our only explicit	0.9181	0.8948	0.7525
Our only non-explicit	0.3458	0.3021	0.3308
Our all senses	0.6136	0.5845	0.5246
Best only explicit	0.9256	0.9022	0.7856
Best only non-explicit	0.4642	0.4091	0.3767
Best all senses	0.6797	0.6434	0.546

Table 2: Overall F_1-measures of discourse relation sense classification evaluated on different relation types on English datasets from our and best competing system of CoNLL 2016 Shared Task (Xue et al., 2016).

that our system performs consistently well on Contingency.Condition, Temporal.Async.Precedence, and Temporal.Async.Succession, but fails on Comparison.Concession, Expansion.Instantiation, and Expansion.Restatement.

Sense	Dev	Test	Blind
Comparison.Concession	0.2000	0.2105	0.0370
Comparison.Contrast	0.7696	0.7690	0.3077
Contingency.Cause.Reason	0.4087	0.5155	0.3556
Contingency.Cause.Result	0.4490	0.4216	0.4110
Contingency.Condition	0.9318	0.8966	0.9811
EntRel	0.5458	0.4523	0.5228
Expansion.Alt	0.9231	0.9091	0.5455
Expansion.Alt.Chosen alt.	0.7692	0.2000	-
Expansion.Conjunction	0.7015	0.6938	0.7432
Expansion.Instantiation	0.2899	0.4496	0.2041
Expansion.Restatement	0.2748	0.2584	0.2378
Temporal.Async.Precedence	0.7812	0.8706	0.8409
Temporal.Async.Succession	0.8211	0.7611	0.8468
Temporal.Synchrony	0.7931	0.6889	0.6034
Overall (micro-average)	0.6136	0.5845	0.5246

Table 3: Per-sense F_1-measures of discourse relation sense classification evaluated on all relations on English datasets.

3.2 Results for Chinese

The training dataset from CDTB for Chinese consists of 455 documents with 2445 discourse relations that can be categorized into 10 different discourse relation senses.

Overall our system performs pretty well on all Chinese datasets (see Table 4) despite not using any external resources or hand-engineered features. Its overall performance is almost consistent across the validation, test (0.7011), and blind (0.7292) datasets, although the last one probably contains a different writing style than CDTB. For only explicit relations our system performs much better on development and test datasets. For non-explicit relations the situation seems to be the opposite. This inconsistencies indicate that the relatively small size of the training dataset does not contain enough information.

Type	Dev	Test	Blind
Our only explicit	0.9351	0.9271	0.7898
Our only non-explicit	0.6667	0.6407	0.7068
Our all senses	0.7206	0.7011	0.7292
Best only explicit	0.9610	0.9634	0.8039
Best only non-explicit	0.7353	0.7242	0.6338
Best all senses	0.7807	0.7701	0.6473

Table 4: Overall F_1-measures of discourse relation sense classification evaluated on different relation types on Chinese datasets from our and best competing system of CoNLL 2016 Shared Task (Xue et al., 2016).

Detailed per-sense analysis on all discourse relations is shown in Table 5. We see that our system performs consistently well on Conjunction, Conditional, and Temporal, but does not perform at all on Alternative, EntRel, and Progression, because of insufficient number of samples.

Sense	Dev	Test	Blind
Alternative	-	-	0.0000
Causation	0.6857	0.4545	0.6748
Conditional	1.0000	0.7500	0.7455
Conjunction	0.8175	0.8228	0.8145
Contrast	0.6957	0.8571	0.6612
EntRel	0.0000	0.0000	0.0000
Expansion	0.5641	0.4628	0.5436
Progression	0.0000	0.0000	0.0000
Purpose	0.8000	0.7857	0.5172
Temporal	1.0000	0.8649	0.7979
Overall (micro-average)	0.7206	0.7011	0.7292

Table 5: Per-sense F_1-measures of discourse relation sense classification evaluated on all relations on Chinese datasets.

4 Conclusion

We have shown that it is possible to implement a shallow discourse relation sense classifier that does not depend on any external sources, hand-engineered features, patterns, and complex fine-

tuned pipelines. Our system consists of two neural network models built from three types of layers and is trained end-to-end. As a consequence it is almost language-agnostic and we have evaluated its performance on the English and Chinese datasets.

References

François Chollet. 2015. Keras. `https://github.com/fchollet/keras`.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv*, pages 1–9.

Xiaojie Jin, Chunyan Xu, Jiashi Feng, Yunchao Wei, Junjun Xiong, and Shuicheng Yan. 2015. Deep Learning with S-shaped Rectified Linear Activation Units.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-Styled End-to-End Discourse Parser. *Nat. Lang. Eng.*, 20(2):151–184.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Nips*, pages 1–9.

Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2014. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Inf. Access Eval. meets Multilinguality, Multimodality, Vis. 5th Int. Conf. CLEF Initiat. (CLEF 14)*, pages 268–299, Berlin Heidelberg New York, sep. Springer.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. *Proc. Sixth Int. Conf. Lang. Resour. Eval.*, pages 2961–2968.

Evgeny Stepanov, Giuseppe Riccardi, and Ali Orkan Bayer. 2015. The UniTN Discourse Parser in CoNLL 2015 Shared Task: Token-level Sequence Labeling with Argument-specific Models. *Proc. Ninet. Conf. Comput. Nat. Lang. Learn. - Shar. Task*, (Dcd):25–31.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proc. Ninet. Conf. Comput. Nat. Lang. Learn. Shar. Task*, pages 17–24.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol T. Rutherford. 2015. The CoNLL-2015 Shared Task on Shallow Discourse Parsing. In *Proc. Ninet. Conf. Comput. Nat. Lang. Learn. Shar. Task*, pages 1–16.

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 *Proc. Twent. Conf. Comput. Nat. Lang. Learn. - Shar. Task*, Berlin, Germany, aug. Association for Computational Linguistics.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style Discourse Annotation of Chinese Text. *Proc. 50th Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.*, (July):69–77.