Do We Really Need All Those Rich Linguistic Features? A Neural Network-Based Approach to Implicit Sense Labeling

Niko Schenk[*], **Christian Chiarcos**[*], **Kathrin Donandt**[*],
Samuel Rönnqvist[*,†], **Evgeny A. Stepanov**[‡] and **Giuseppe Riccardi**[‡]
[*]Applied Computational Linguistics Lab, Goethe University, Frankfurt am Main, Germany
[†]Turku Centre for Computer Science, TUCS, Åbo Akademi University, Turku, Finland
[‡]Signals and Interactive Systems Lab, DISI, University of Trento, Italy
{schenk,chiarcos,donandt}@informatik.uni-frankfurt.de,
sronnqvi@abo.fi, {evgeny.stepanov,giuseppe.riccardi}@unitn.it

Abstract

We describe our contribution to the CoNLL 2016 Shared Task on shallow discourse parsing.[1] Our system extends the two best parsers from previous year's competition by integration of a novel *implicit* sense labeling component. It is grounded on a highly generic, language-independent feedforward neural network architecture incorporating weighted word embeddings for argument spans which obviates the need for (traditional) hand-crafted features. Despite its simplicity, our system overall outperforms all results from 2015 on 5 out of 6 evaluation sets for English and achieves an absolute improvement in F_1-score of 3.2% on the PDTB test section for non-explicit sense classification.

1 Introduction

Text comprehension is an essential part of Natural Language Understanding and requires capabilities beyond capturing the lexical semantics of individual words or phrases. In order to understand how meaning is established, altered and transferred across words and sentences, a model is needed to account for contextual information as a semantically coherent representation of the logical *discourse structure* of a text. Different formalisms and frameworks have been proposed to realize this assumption (Mann and Thompson, 1988; Lascarides and Asher, 1993; Webber, 2004).

In a more applied NLP context, *shallow discourse parsing* (SDP) aims at automatically detecting relevant discourse units and to label the relations that hold between them. Unlike *deep discourse parsing*, a stringent logical formalization or the establishment of a global data structure, for instance, a tree, is not required.

With the release of the Penn Discourse Treebank (Prasad et al., 2008, PDTB) and the Chinese Discourse Treebank (Zhou and Xue, 2012, CDTB), annotated training data for SDP has become available and, as a consequence, the field has considerably attracted researchers from the NLP and IR community. Informally, the PDTB annotation scheme describes a discourse unit as a syntactically motivated character span in the text, augmented with relations pointing from the second argument (*Arg2*, prototypically, a discourse unit associated with an explicit discourse marker) to its antecedent, i.e., the discourse unit *Arg1*. Relations are labeled with a relation type (its *sense*) and the associated discourse marker (either as found in the text or as inferred by the annotator). PDTB distinguishes *explicit* and *implicit* relations depending on whether such a connector or cue phrase (e.g., *because*) is present, or not.[2] As an illustrative example without such a marker, consider the following two adjacent sentences from the PDTB:

Arg1: *The real culprits are computer makers such as IBM that have jumped the gun to unveil 486-based products.*
Arg2: *The reason this is getting so much visibility is that some started shipping and announced early availability.*

In this *implicit* relation, *Arg1* and *Arg2* are directly related. The discourse relation type is *Expansion.Restatement*—one out of roughly twenty finegrained tags marking the sense relation

[1] http://www.cs.brandeis.edu/~clp/conll16st
Our parser code is available at: https://github.com/acoli-repo/shallow-discourse-parser

[2] The set of relation types is completed by alternative lexicalization (*AltLex*, discourse marker rephrased), entity relation (*EntRel*, i.e., anaphoric coherence), resp. the absence of any relation (*NoRel*).

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 41–49,
Berlin, Germany, August 7-12, 2016.

between any given argument pair in the PDTB.

Our Contribution: We participate in the CoNLL 2016 Shared Task on SDP (Xue et al., 2016; Potthast et al., 2014) and propose a novel, neural network-based approach for implicit sense labeling. Its system architecture is modular, highly generic and mostly language-independent, by leveraging the full power of pre-trained word embeddings for the SDP sense classification task. Our parser performs well on both English and Chinese data and is highly competitive with the state-of-the-art, though does not require manual feature engineering as employed in most prior works on implicit SDP, but rather relies extensively on features learned from data.

2 Related Work

Most of the literature on automated discourse parsing has focused on specialized subtasks such as:

1. **Argument identification**
   (Ghosh et al., 2012; Kong et al., 2014)
2. **Explicit sense classification**
   (Pitler and Nenkova, 2009)
3. **Implicit sense classification**
   (Marcu and Echihabi, 2002; Pitler et al., 2009; Lin et al., 2009; Zhou et al., 2010; Park and Cardie, 2012; Biran and McKeown, 2013; Rutherford and Xue, 2014)

A minimal requirement for any full-fledged end-to-end discourse parser is to integrate at least these three processes into a sequential pipeline. However, until recently, only a handful of such parsers have existed (Lin et al., 2014; Biran and McKeown, 2015; duVerle and Prendinger, 2009; Feng and Hirst, 2012). It has been enormously difficult to evaluate the performance of these systems among themselves, and also to compare the efficiency of their individual components with other competing methods, as i.) those systems rely on different theories of discourse, e.g., PDTB or RST; and ii) different (sub)modules involve custom settings, feature- and tool-specific parameters, (esp. for the most challenging task of *implicit sense labeling*). Furthermore, iii) most previous works are not directly comparable in terms of overall accuracies as their underlying evaluation data suffers from inconsistent label sizes among studies (e.g., full sense inventory vs. simplified 1- or 2-level classes, cf. Huang and Chen (2011)).

Fortunately, with the first edition of the shared task on SDP, Xue et al. (2015) had established a *unified framework* and had made an independent evaluation possible. The best performing participating systems – most notably those by Wang and Lan (2015) and Stepanov et al. (2015) – have re-implemented the well-established techniques, for example the one by Lin et al. (2014).

2.1 Deep Learning Approaches to SDP

In last year's shared task, first implementations on *deep learning* have seen a surge of interest: Wang et al. (2015) and Okita et al. (2015) proposed a recurrent neural network for argument identification and a paragraph vector model for sense classification. Distributed representations for both arguments were obtained by vector concatenation of embeddings.

An earlier attempt in a similar direction of *representation learning* (Bengio et al., 2013) has been made by Ji and Eisenstein (2014). The authors demonstrated successfully how to discriminatively learn a latent, low-dimensional feature representation for RST-style discourse parsing, which has the benefit of capturing the underlying meaning of elementary discourse units without suffering from data sparsity of the originally high dimensional input data.

Closely related, Li et al. (2014) introduced a recursive neural network for discourse parsing which jointly models distributed representations for sentences based on words and syntactic information. The approach is motivated by Socher et al. (2013) and models the discourse unit's root embedding to represent the whole discourse unit which is being obtained from its parts by an iterative process. Their system is made up of a binary structure classifier and a multi-class relation classifier and achieves similar performance compared to Ji and Eisenstein (2014).

Very recently, Liu et al. (2016) and Zhang et al. (2015) have successfully applied convolutional neural networks to model implicit relations within the PDTB-framework. Along these lines and inspired by the work in Weiss (2015), we also see great potential in the use of neural network-based techniques to SDP. Similarly, our approach trains a modular component for shallow discourse parsing which incorporates distributed word representations for argument spans by abstraction from surface-level (token) information. Crucially, our

approach substitutes the traditional sparse and hand-crafted features from the literature to account for a minimalist, but at the same time, general (latent) representation of the discourse units. In the next sections, we elaborate on our novel neural network-based approach for implicit sense labeling and how it is fit into the overall system architecture of the parser.

3 A Neural Sense Labeler for Implicit and Entity Relations

We construct a neural network-based module for the classification of senses for both implicit and entity (*EntRel*) relations.[3] As a very general and highly data-driven approach to modeling discourse relations, our classifier incorporates *only* word embeddings and basic syntactic dependency information. Also, in order to keep the setup easily adaptable to new data and other languages, we avoid the use of very specific and costly hand-crafted features (such as sentiment polarities, word-pair features, cue phrases, modality, production rules, highly specific semantic information from external ontologies such as VerbNet, etc.), which has been the main focus in traditional approaches to SDP (Huang and Chen, 2011; Park and Cardie, 2012; Feng and Hirst, 2012). Instead, we substitute (sparse) tokens in the argument spans, with dense, distributed representations, i.e. word embeddings, as the main source of information for the sense classification component. Closely related, Zhang et al. (2015) have explored a similar approach of constructing argument vectors by applying a set of aggregation functions on their token vectors, however, without the use of additional (syntactic) information, while embedding their vectors into a single-layer neural network only.

In our experiments, we used the pre-trained *GoogleNews* vectors (for English) and the *Gigaword*-induced vectors (for Chinese) provided by the shared task as a starting point.[4] We further trained the word vectors on the raw Wall Street Journal texts, thus tuning the embeddings toward the data at hand, with the goal of considerably improving their predictive power in the sense classification task. Specifically, the pre-trained vectors of size 300 were updated by the skip-gram method (Mikolov et al., 2013)[5] in multiple passes over the Newswire texts with decreasing learning rate. This procedure is supposed to improve the quality of the embeddings and also their coverage.

Our new word vector model provides general vector representations for each token in the two argument spans[6], which forms the basis for producing compositional vectors to represent the two spans. Compositional vectors that introduce a fixed-length representation of a variable-length span of tokens are practical features for feedforward neural networks. Thus, we may combine the token vectors of each span by simply averaging vectors, or – following Mitchell and Lapata (2008) – by calculating an aggregated argument vector $\vec{v'}$:

$$\vec{v'}(j) = \frac{1}{k(j)} \sum_{i=1}^{k(j)} V(j)_i + \prod_{i=1}^{k(j)} V(j)_i \qquad (1)$$

for arguments $j \in \{1, 2\}$, where $k(j) = |t(j)|$ defines their lengths in the number of tokens and $\prod$ applies the pointwise product $\odot$ over the token vectors in $V(j)$.

Both procedures produce rather simple argument representations that do not account for word order variation or any other sentence structure information, yet they serve as decent features for discourse parsing and other related tasks. By introducing pointwise multiplication of the token vectors, the elements that represent assumed independent, latent semantic dimensions are not merely lumped together across vectors, but are allowed to scale according to their mutual relevance.[7]

Improving upon the compositional representation produced by Equation 1, we incorporate additional syntactic dependency information: for each token in an argument span, we calculate the depth d from the corresponding sentence's root node and weight the token vector by $\frac{1}{2^d}$ before applying the

[3] The reason to combine both relation types has been a design decision as *EntRels* are very similar to implicit relations and are also missing a connective. *AltLex* relations seemed too few to have any statistical impact on the performance of our experiments and have been ignored altogether.

[4] `http://www.cs.brandeis.edu/~clp/conll16st/dataset.html`

[5] We found window size of 8 and min term count = 3 to be optimal. Neural networks were trained using the *gensim* package: `http://radimrehurek.com/gensim/`.

[6] We ignore unknown tokens for which no vectors exist.

[7] In our experiments, Equation 1 outperformed simpler strategies of either average or multiplication alone. This also indicates that it is beneficial to not completely suppress dimensions with near-zero values for single tokens.

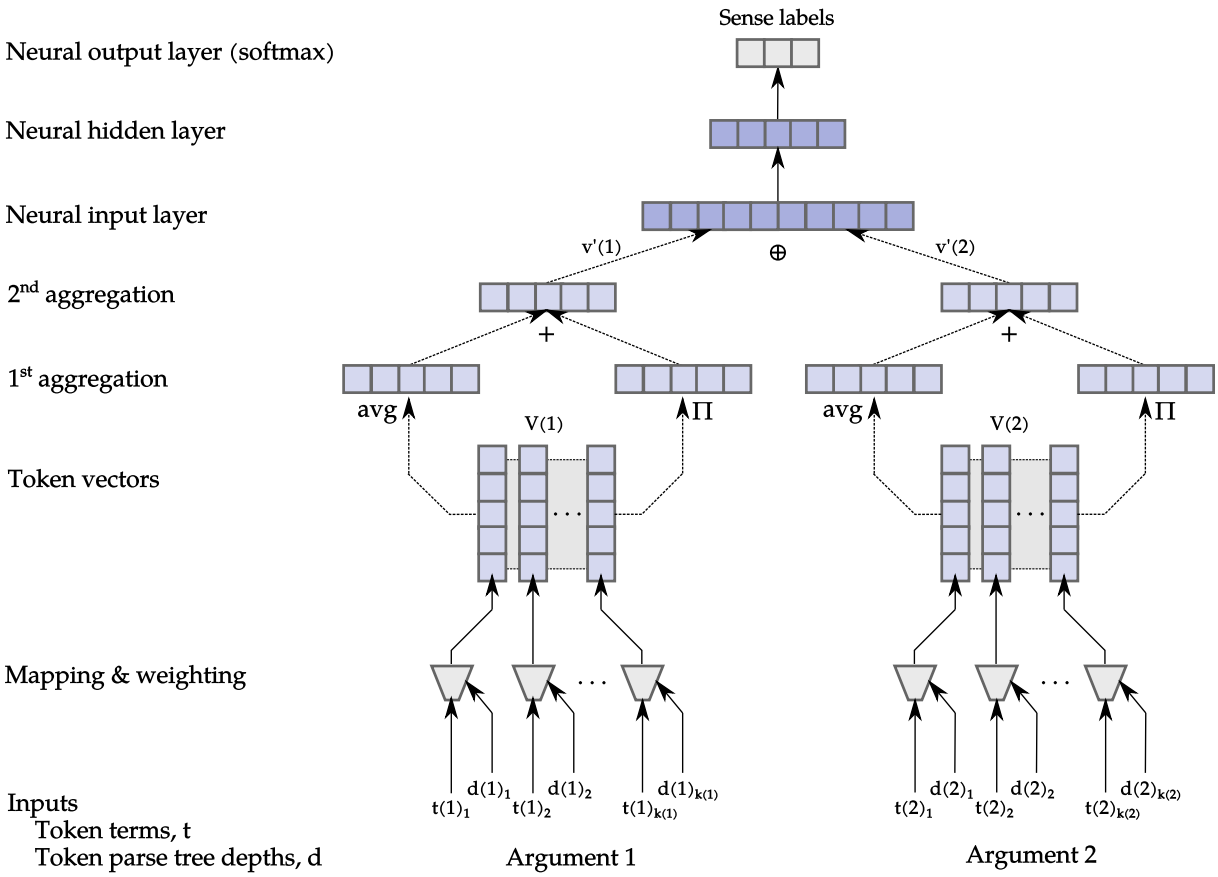

Figure 1: The feature construction process from argument spans (light blue) and neural architecture (dark blue) for implicit sense classification (incl. *EntRel*) . Dotted lines represent pointwise vector operations.

aggregating operators.[8]

The bottom of Figure 1 illustrates the first step of the process, i.e. mapping tokens to their corresponding vectors based on the updated word vector model, as well as the token depth weighting. Secondly, the aggregation operators are applied, i.e., the sum ($+$) of the pointwise product ($\prod/\odot$) and average (*avg*) of the vectors. Finally, the compositional vectors for each of the arguments are concatenated ($\oplus$) and serve as input to a feedforward neural network.

Given the composed argument vectors, we set up a network with one hidden layer and a softmax output layer to classify among 20 implicit senses for English and 9 for Chinese, plus an additional *EntRel* label. Other relations, such as *AltLex*, are not modeled. We train the network using Nesterov's Accelerated Gradient (Nesterov, 1983) and optimized all hyper-parameters on the development set. Best results were achieved with *rectified linear activation with learnable leak rate and gain (lgrelu)*, 40-60 hidden nodes and weight decay and hidden node regularization of 0.0001.[9]

4 The Competition Tasks & Pipelines

We participate in the *closed track* of the shared task, specifically in both *full* and *supplementary tasks* (*sense-only*) on English and Chinese texts. Full tasks require a participant's system to identify argument pairs and to label the sense relation that holds between them. In each supplementary task, gold arguments are provided so that the performance of sense labeling does not suffer from error propagation due to incorrectly detected argument spans.

We combine different *existent* modules to address the specific settings and classification needs of both full and supplementary tasks for both lan-

[8]Tokens that are missing in the parse tree, such as punctuation symbols, are weighted by 0.25, in our optimal setting.

[9]The learning rate was set to 0.0001. Momentum of 0.35-0.6 and 60 hidden nodes performed well for the English tasks, and momentum of 0.85 and 40 hidden nodes for Chinese (with fewer output nodes). Good results were also obtained by *Parametric Rectified Linear Unit (prelu)* activation, as well as the combination of larger hidden layer and stronger regularization (e.g., L1 regularization of 0.1 on 100 nodes).

guages. The modules and their combination with our implicit neural sense classifier will be outlined in the following sections.

4.1 English Full Task Pipeline (EFTP)

For the full task, we exploit the high-quality argument extraction modules of the two best-performing systems by Wang and Lan (2015, W&L) and Stepanov et al. (2015) from last year's competition (re-using their original implementations): Specifically, we initially run both systems for all *explicit* relations only, and keep those predicted arguments and sense labels – from either of the two systems – which maximize F_1-score on the development set. With this simple heuristic, we hope to improve upon the best results from W&L, as, for instance, Stepanov et al. (2015) perform particularly well on all temporal relations, while W&L's tool handles the majority of other senses well.

For all implicit and *EntRel* relations, we keep the exact argument spans obtained from the W&L system and reject all sense labels. In a second step, we *re-classify* all these implicit relations by our neural net-based architecture described in Section 3 given only the tokens and their dependencies in both argument spans. Finally, we merge all combined explicit and re-classified implicit relations into the final set for evaluation.

4.2 English Supplementary Task Pipeline (ESTP)

We make use of the system by Stepanov et al. (2015) to label all *explicit* relation senses, and classify all other relations with an empty token list for connectors (i.e., implicit and *EntRels*) by our neural network architecture from Section 3.

4.3 Chinese Full Task Pipeline (CFTP)

Since for the Chinese full task no reusable argument extraction tools were available, we have set up a minimalist (baseline) implementation whose individual steps we sketch briefly:

1. **Connective detection** is realized by means of a sequence labeling/CRF model.[10] Features are unigram and bigram information from the tokens, their parts-of-speech, dependency head, dependency chain, whether the token is found as a connector in the training set, and its relative position within the sentence.
2. **Argument extraction** is based on the output of predicted connectives for both inter- and intra-sentence relations. As an additional feature, we found the IOB chain for the syntactic path of a token to be useful.[11]
3. We heuristically **post-process** the CRF-labeled argument tokens in order to assign connectors to same-sentence or separate-sentence *Arg1* and *Arg2* spans.
4. The so-obtained **explicit argument pairs** are sense labeled by a (linear-kernel) SVM classifier[12] with the connector word as the only feature, following the minimalist setting in Chiarcos and Schenk (2015).
5. As **implicit relations** we consider *all inter-sentential relations* which are not already part of an explicit relation. Same-sentence relations are ignored altogether.

4.4 Chinese Supplementary Task Pipeline (CSTP)

For the provided argument pairs, we label *explicit* relations (i.e. those containing a non-empty connector) by the SVM classifier which has been trained using only a single feature – the connector token. For all other relations, we again employ our neural network-based strategy described in Section 3. The overall architecture is exactly the same as for the English subtask; only the (hyper)parameters have been updated in accordance with the Chinese training data.

5 Evaluation

5.1 English Full Task

Table 1 shows the performance of our full-task pipeline (EFTP) which integrates our novel feed-forward neural network architecture for implicit sense labeling. The figures suggest that our minimalist approach is highly competitive and can even outperform the best results from last year's competition in terms of F_1-scores on two out of three evaluation sets (cf. last *implicit* column).

Overall, with the integration of the combined systems by W&L and Stepanov et al. (2015), we can improve upon the state-of-the-art by an absolute increase in F_1-score of 0.5% on the blind test

[10] https://taku910.github.io/crfpp/

[11] This information was generated using the script from http://ilk.uvt.nl/team/sabine/chunklink/chunklink_2-2-2000_for_conll.pl

[12] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

set– which is marginal but only due to the fruitful re-classification of the already-provided (and therefore fixed) argument spans.

Measured on the development set, we found that the *dependency depth weighting* contributes to an absolute improvement in accuracy of 1.5% for non-explicit relations.

set	system	overall	explicit	*implicit*
dev	W&L	37.84	48.16	28.70
	EFTP	**40.21**	**50.87**	**30.99**
test	W&L	29.69	39.96	**20.74**
	EFTP	**29.78**	**40.44**	20.60
blind	W&L	24.00	30.38	18.78
	EFTP	**24.47**	**30.74**	**19.63**

Table 1: English full task F_1-scores.

5.2 English Supplementary Task

Without error propagation from argument identification, and with the gold arguments provided in the evaluation sets, the performance of our implicit sense labeling component is even better; cf. Table 2: on both PDTB evaluation sets F_1-scores increase by 2.7% and 3.16% (absolute) and by 6.32% and up to **9.17%** (relative) on the development and test section, respectively.

Strikingly, however, the prediction quality on the blind test set is worse than expected. We assume that this is partly due to the (slightly) heterogeneous content of the annotated *Wikinews*, as opposed to the original Penn Discourse Treebank data on which our system performs extraordinarily well.

set	system	overall	explicit	*implicit*
dev	W&L	65.11	90.00	42.72
	ESTP	**66.90**	**91.35**	**45.42**
test	W&L	61.27	**90.79**	34.45
	ESTP	**62.64**	90.13	**37.61**
blind	W&L	**54.76**	**76.44**	**36.29**
	ESTP	52.32	76.40	31.85

Table 2: English sense-only task F_1-scores.

5.3 Chinese Full Task

This year's edition of the shared task has been the first to address shallow discourse parsing for Chinese Newswire texts. Given no prior (directly comparable) results on Chinese SDP so far, we simply report the performance of our system on all evaluation sets in Table 3.

set	system	overall	explicit	*implicit*
dev	**CFTP**	22.16	17.45	**22.67**
test	**CFTP**	**24.21**	**28.73**	22.26
blind	**CFTP**	12.90	18.56	10.80

Table 3: Chinese full task F_1-scores.

5.4 Chinese Supplementary Task

A final evaluation has been concerned with the sense-only labeling of gold-provided arguments for Chinese. We want to point out that the neural network architecture for implicit relations (with 70.59% F_1-score on the dev set, cf. Table 4) has beaten all our other experiments: In particular, we have conducted an SVM setup in which we employed the traditional word-pair features substituted by Brown clusters 3200 (65.12%), and special additive Arg1/Arg2 combinations of word embeddings – yielding only 62.8% which equals the majority class baseline indicating no predictive power for any given kernel type.

set	system	overall	explicit	*implicit*
dev	**CSTP**	75.72	96.10	70.59
test	**CSTP**	**77.01**	**96.34**	**71.87**
blind	**CSTP**	63.73	80.39	57.59

Table 4: Chinese sense-only task F_1-scores.

6 Conclusion

In the context of the CoNLL 2016 Shared Task on shallow discourse parsing, we have described our participating system and its architecture. Specifically, we introduced a novel feedforward neural network-based component for implicit sense labeling whose only source of information are pre-trained word embeddings and syntactic dependencies. Its highly generic and extremely simple design is the main advantage of this module. It has proven to be language-independent, easy to tune and optimize and does not require the use of handcrafted – rich – linguistic features.

Still its performance is highly competitive with the state-of-the-art on implicit sense labeling and builds a solid groundwork for future extensions.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August.

Or Biran and Kathleen McKeown. 2013. Aggregated Word Pair Features for Implicit Discourse Relation Disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 69–73.

Or Biran and Kathleen McKeown. 2015. PDTB Discourse Parsing as a Tagging Task: The Two Taggers Approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 96–104, Prague, Czech Republic, September. Association for Computational Linguistics.

Christian Chiarcos and Niko Schenk. 2015. A Minimalist Approach to Shallow Discourse Parsing and Implicit Relation Recognition. In *Proceedings of the 19th Conference on Computational Natural Language Learning: Shared Task, CoNLL 2015, Beijing, China, July 30-31, 2015*, pages 42–49.

David A. duVerle and Helmut Prendinger. 2009. A Novel Discourse Parser Based on Support Vector Machine Classification. 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*, ACL '09, pages 665–673, Stroudsburg, PA, USA. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level Discourse Parsing with Rich Linguistic Features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 60–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sucheta Ghosh, Giuseppe Riccardi, and Richard Johansson. 2012. Global Features for Shallow Discourse Parsing. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 150–159.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese Discourse Relation Recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1442–1446, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation Learning for Text-level Discourse Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland, June. Association for Computational Linguistics.

Fang Kong, Tou Hwee Ng, and Guodong Zhou. 2014. A Constituent-Based Approach to Argument Labeling with Joint Inference in Discourse Parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 68–77. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 1993. Temporal Interpretation, Discourse Relations and Commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive Deep Models for Discourse Parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha, Qatar, October. Association for Computational Linguistics.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 343–351, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184, 4.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit Discourse Relation Classification via Multi-Task Neural Networks. *CoRR*, abs/1603.02776.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu and Abdessamad Echihabi. 2002. An Unsupervised Approach to Recognizing Discourse Relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 368–375, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at International Conference on Learning Representations*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based Models of Semantic Composition. In *Proceedings of Association for Computational Linguistics*, pages 236–244.

Yurii Nesterov. 1983. A method of solving a convex programming problem with convergence rate O (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376.

Tsuyoshi Okita, Longyue Wang, and Qun Liu. 2015. The DCU Discourse Parser: A Sense Classification Task. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 71–77, Beijing, China, July. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2012. Improving Implicit Discourse Relation Recognition Through Feature Set Optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 108–112, Seoul, South Korea, July. Association for Computational Linguistics, Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using Syntax to Disambiguate Explicit Discourse Connectives in Text. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers*, pages 13–16.

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*, ACL '09, pages 683–691, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Berlin Heidelberg New York, September. Springer.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *In Proceedings of LREC*.

Attapol 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*, pages 645–654. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Stroudsburg, PA, October. Association for Computational Linguistics.

Evgeny Stepanov, Giuseppe Riccardi, and Orkan Ali Bayer. 2015. The UniTN Discourse Parser in CoNLL 2015 Shared Task: Token-level Sequence Labeling with Argument-specific Models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 25–31. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Longyue Wang, Chris Hokamp, Tsuyoshi Okita, Xiaojun Zhang, and Qun Liu. 2015. The DCU Discourse Parser for Connective, Argument Identification and Explicit Sense Classification. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 89–94. Association for Computational Linguistics.

Bonnie L. Webber. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science*, 28(5):751–779.

Gregor Weiss. 2015. Learning Representations for Text-level Discourse Parsing. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 16–21, 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*, Beijing, China.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Bonnie Webber, Attapol Rutherford, Chuan Wang, and Hongmin Wang. 2016. The CoNLL-2016 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task*, Berlin, Germany, August. Association for Computational Linguistics.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2230–2235.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style Discourse Annotation of Chinese Text. In *Proceedings of the 50th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 69–77, Jeju Island, Korea, July. Association for Computational Linguistics.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting Discourse Connectives for Implicit Discourse Relation Recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 1507–1514, Stroudsburg, PA, USA. Association for Computational Linguistics.