Learning representations for text-level discourse parsing

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

Abstract

In the proposed doctoral work we will design an end-to-end approach for the challenging NLP task of text-level discourse parsing. Instead of depending on mostly hand-engineered sparse features and independent components for each subtask, we propose a unified approach completely based on deep learning architectures. To train more expressive representations that capture communicative functions and semantic roles of discourse units and relations between them, we will jointly learn all discourse parsing subtasks at different layers of our architecture and share their intermediate representations. By combining unsupervised training of word embeddings with our layer-wise multi-task learning of higher representations we hope to reach or even surpass performance of current state-of-the-art methods on annotated English corpora.

1 Introduction

Modern algorithms for natural language processing (NLP) are based on statistical machine learning and require a computationally convenient representation of input data. Unfortunately real-world plain text is usually represented as an unstructured sequence of words with complex relations between them. Therefore it is extremely important to discover good representations in the form of informative text features.

In NLP such features are almost always hand-engineered sparse features and require expensive human labor and expert knowledge to construct. They are usually based on lexicons or features extracted by other NLP subtasks and have the form of hand-engineered extraction rules, regular expressions, lemmatization, part-of-speech (POS) tags, positions or lengths of arguments, tense forms, syntactic parse trees, and similar. Although such features are specific for a given language, domain, and task, they work well enough for simple NLP tasks, like named entity recognition or POS tagging. Nevertheless, the ability to learn text features and representations automatically would have a lot of potential to improve state-of-the-art performance on more challenging NLP tasks, such as text-level discourse parsing. This may even be more important for languages where progress in NLP is still lacking.

Variants of deep learning architectures have been shown to provide a different approach to learning in which latent features are automatically learned as distributed dense vectors. They managed to represent meaningful relations with word (Collobert, 2011), POS and dependency tag (Chen and Manning, 2014), sentence (Guo and Diab, 2012), and document (Socher et al., 2012) embeddings and achieved surprising results for a number of NLP tasks. It has been shown that both unsupervised pre-training (Hinton et al., 2006) and multi-task learning (Collobert and Weston, 2008) significantly improve their performance in the absence of hand-engineered features. This makes them especially interesting for the problem of text-level discourse parsing.

2 Text-level discourse parsing

In natural language, a piece of text meant to communicate specific information, function, or knowledge (clauses, sentences, or even paragraphs) is called a discourse. They are often understood only in relation to other discourse units (at any level of grouping) and their combination creates a joint meaning larger than individual unit's meaning alone (Mann and Thompson, 1988).

Discourse parsing is the task of determining how these units are related to each other (like in Figure 1) and plays a central role in a num-

ber of high-impact natural language processing (NLP) applications, including text summarization, sentence compression, sentiment analysis, and question-answering. For analyzing different perspectives of discourse analysis researchers proposed a number of theoretical frameworks and released annotated corpora, such as RST Discourse Treebank (RST-DT) (Carlson et al., 2003) and Penn Discourse Treebank (PDTB) (Prasad et al., 2008). Both of these decompose discourse parsing into a few subtasks and, like in most of NLP, their success depends on expert knowledge of each subtask and hand-engineering of more powerful features (Feng and Hirst, 2012; Lin et al., 2014), representations, and heuristics (Joty et al., 2013; Prasad et al., 2010).

Despite recent progress in automatic discourse segmentation and sentence-level parsing (Fisher and Roark, 2007; Joty et al., 2012; Soricut and Marcu, 2003), text-level discourse parsing remains a significant challenge (Feng and Hirst, 2012; Ji and Eisenstein, 2014; Lin et al., 2014). Traditional hand-engineering approaches unfortunately seem to be insufficient, as discourses and relations between them do not follow any strict grammar or obvious rules.

Two main theoretical frameworks with English corpus have been proposed to capture different rhetorical characteristics, and serve different applications.

The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) is currently the largest discourse-annotated corpus, consisting of 2159 articles from Wall Street Journal. It strives to maintain a theory-neutral approach by adopting the predicate-argument view and independence of discourse relations. In it either explicitly or implicitly given discourse connectives, such as coordinating conjunction (e.g. "and", "but"), subordinating conjunction (e.g. "if", "because"), or discourse adverbial (e.g. "however", "also"), combine pairs of discourse arguments into relations. For PDTB-style discourse parsing, extracting argument spans seems to be the most difficult subtask (Lin et al., 2014), resulting in the best overall performance of only 34.80% in F_1-measure (Kong et al., 2014).

The RST Discourse Treebank (RST-DT) (Carlson et al., 2003) follows the theoretical framework of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). It contains 385 annotated documents from the Wall Street Journal with 18 high-level categories and 110 fine-grained relations. Any coherent text can be represented as a RST discourse tree structure (like in Figure 1) whose leaves are minimal non-overlapping text spans called elementary discourse units. Adjacent nodes are joined depending on their discourse relations to form a tree. In a mono-nuclear discourse relation one of the text spans is the nucleus, which is more salient than the satellite, while in a multi-nuclear relation all text spans are equally important for interpretation. Performance of RST-style discourse parsing is evaluated based on their ability to locate spans of text that serve as arguments (best 85.7% in F_1-measure (Feng and Hirst, 2012)), identify which of the arguments is the nucleus (best 71.1% in F_1-measure (Ji and Eisenstein, 2014)), and tag the sense and location of discourse relations (best 61.6% in F_1-measure (Ji and Eisenstein, 2014)).

3 Related work

Early work on linguistic and computational discourse analysis produced several theoretical frameworks and one of the most influential is Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In order to automatically build a hierarchical structure of a text, first approaches (Marcu, 2000) relied mainly on discourse markers, hand-engineered rules, and heuristics. Learning-based approaches were first applied to identify within-sentence discourse relations (Soricut and Marcu, 2003), and only later to cross-sentence text-level relations (Baldridge and Lascarides, 2005). They largely focused on lexical, syntactic, and structural features, but the close relationship between discourse structure and semantic meaning suggests that this may not be sufficient (Prasad et al., 2008; Subba and Di Eugenio, 2009). Further work on discourse parsing focused first on having a binary classifier for determining whether two adjacent discourse units should be merged, followed by a multi-class classifier for determining which discourse relation should be assigned to the new subtree (DuVerle and Prendinger, 2009). Improved results (Feng and Hirst, 2012) have been achieved by incorporating rich linguistic features (Hernault et al., 2010), including lexical semantics, and specific discourse production rules (Lin et al., 2009). An alternative approach is based on jointly performing detection and classification in a bottom-

- [The dollar finished lower yesterday,] e_1 [after another session on Wall Street.] e_2
- [Concern about the volatile U.S. stock market had faded in recent sessions,] e_3 [and traders let the dollar languish in a narrow range until tomorrow,] e_4 [when the preliminary report on U.S. gross national product is released.] e_5
- [But movements in the Dow Jones Industrial Average yesterday put Wall Street back in the spotlight] e_6 [and inspired participants to bid the U.S. unit lower.] e_7

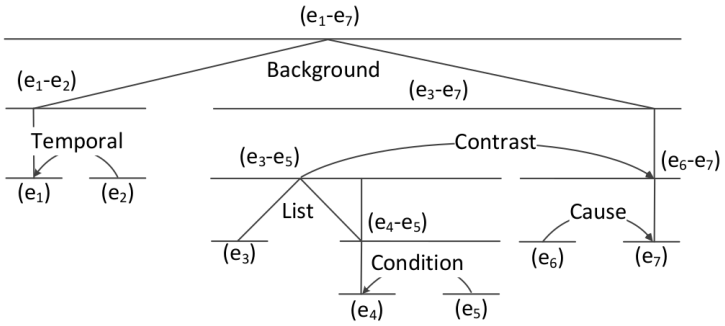

Figure 1: An example of seven elementary discourse units (e_1-e_7), and (mono- or multi-nuclear) relations between them in an RST discourse tree representation (Feng et al., 2014).

up fashion while distinguishing within-sentence and cross-sentence relations (Joty et al., 2013) and improved with discriminative reranking of discourse trees using tree kernels (Joty and Moschitti, 2014). It has been shown that constituent- and dependency-based syntax and features based on coreference links improve performance (Surdeanu et al., 2015). The first PDTB-style end-to-end discourse parser (Lin et al., 2014) uses a connective list to identify explicit candidates, followed by simple features and parse trees to extract arguments and identify discourse relations. Classifying implicit discourse relations can be improved by combining distributed representations of parse trees with coreferent entity mentions (Ji and Eisenstein, 2015). Extracting discourse arguments has been attempted by using classic linear word tagging with conditional random fields and global features (Ghosh et al., 2012), identifying nodes in constituent subtrees (Lin et al., 2014), and hybrid merging and pruning of parse trees with integer linear programming (Kong et al., 2014).

Deep learning architectures consist of multiple layers of simple learning blocks stacked on each other and, when well trained, tend to do a better job at disentangling the underlying factors of variation. Beginning with raw data, its representation is transformed into increasingly higher and more abstract forms in each layer, until the final low-dimensional features or representation useful for a given task is reached. Their success is possible with breakthroughs and improvements in training techniques (like AdaGrad or Adam optimization, rectifier function, dropout regularization) and with initialization using unsupervised pre-training (Hinton et al., 2006; Collobert, 2011) on massive datasets (such as Wikipedia or Wall Street Journal). Pre-training helps deep networks to develop natural abstractions and combined with multi-task learning (Collobert and Weston, 2008) it can significantly improve their performance in the absence of hand-engineered features.

Classic feed-forward architectures are inappropriate for processing text documents, because of their variable length and natural representation as a sequence of words. One approach to solve this is to specify a transition-based processing mechanism (Chen and Manning, 2014; Ji and Eisenstein, 2014) and train a neural network classifier to make parsing decisions. Recurrent neural networks (RNNs) (Elman, 1990) or their generalization, recursive neural networks (Goller and Küchler, 1996), represent a more direct approach by recursively applying the same set of weights over the sequence (temporal dimension) or structure (tree-based). Li et al. (Li et al., 2015) have recently

showed that only some NLP tasks benefit from recursive models applied on syntactic parse trees and recurrent models seem to be sufficient for discourse parsing. By stacking multiple hidden layers into a deep RNN makes them represent a temporal hierarchy with multiple layers operating at different time scales (Hermans and Schrauwen, 2013). Learning to store information over extended time intervals has been achieved with long short-term memory (Hochreiter and Schmidhuber, 1997), time delay neural network (Waibel et al., 1989), or neural Turing machines (Graves et al., 2014). Bidirectional variants of these models can incorporate information from preceding as well as following tokens (Schuster and Paliwal, 1997). Recursive neural networks have also been shown to support different task-specific representations, such as matrix-vector representation of words (Socher et al., 2012) or recurrent neural tensor networks (Socher et al., 2013). For our discourse parsing task such deeper models, that can learn abstract representations on different time scales, might better model the discourse relations between input vectors and (hopefully) capture their communicative functions and semantic meaning.

A few initial attempts of applying representation learning to our task have already shown substantial performance improvements over previous state-of-the-art. Ji and Eisenstein (Ji and Eisenstein, 2014) implement a shift-reduce discourse parser on top of given RST-style discourse units to simultaneously learn parsing and a discourse-driven projection of features using support vector machines with gradient-based updates. Li et al. (Li, 2014) produce a distributed representation of RST-style discourse units using recursive convolution on sentence parse trees and apply a classifier to determine relations between them. Ji and Eisenstein (Ji and Eisenstein, 2014) also improved classification of PDTB-style implicit discourse relations by combining distributed representations of parse trees with coreferent entity mentions.

4 Contribution to science

Because text-level discourse parsing is an important, yet still challenging NLP task, it is the focus of our doctoral dissertation.

Method for text-level discourse parsing. Instead of depending on mostly hand-engineered sparse features and independent separately-developed components for each subtask, we propose a unified end-to-end approach for text-level discourse parsing completely based on deep learning architectures. First each of the discourse parsing subtasks, such as argument boundary detection, labeling, discourse relation identification and sense classification, need to be formulated in terms of RNNs and similar derivable learning architectures. To benefit from their ability to learn intermediate representations they will be partially stacked on top of each order, such that the last but one layer (i.e. output layer) for each subtask is shared with other subtasks. By placing increasingly more difficult subtasks at different layers in one deep architecture, they can benefit from each others intermediate representations, improve robustness and training speed. Figure 2 further combines unsupervised training of word embeddings with our layer-wise multi-task learning of higher representations and illustrates our goal of a unified end-to-end approach for text-level discourse parsing utilizing different layers of representations.

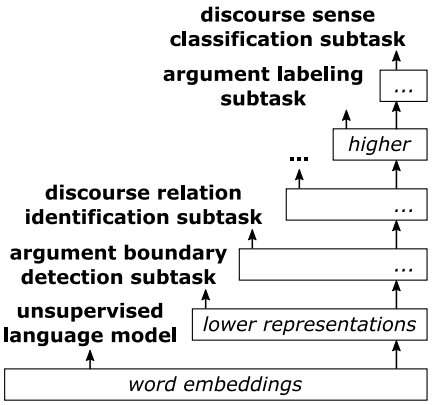

Figure 2: Illustration of our unified end-to-end approach for text-level discourse parsing with layer-wise multi-task learning of higher representations.

5 Work plan

To accomplish this we will, on one hand, need to find the best deep learning models for each of the discourse parsing subtasks, suitable architecture, activation functions, and figure out how to adapt them to operate on sequential data and with each other. This includes analyzing deep learning architectures, identifying their strengths, useful components, and their suitability for our NLP task.

Afterwards combine them into one unified deep learning architecture with shared intermediate rep-

resentations and unsupervised training of word embeddings. Developing a prototype for shallow discourse parsing will open the door for find the best initialization procedures, training functions, learning rates, and similar. Shallow PDTB-style discourse parsing is also a challenge on this years CoNLL 2015 conference, where adjacent text spans are not necessarily connected with discourse relations to form a tree.

Additionally we will experiment with new and more expressive representations and structures (like neural tensor networks) that could capture communicative functions and semantic roles of discourse units and relations between them.

Even though our method could be applied to any plain text, we plan on evaluating it on standard annotated English corpora. After applying our approach on at lest one of the corpora, we intend to qualitatively analyze the identified discourse units and relations between them to gain insights about its strengths and weaknesses. On the other hand, the dataset will allow us to also quantitatively compare its performance to current state-of-the-art methods. The procedure for our method will begin by pre-training the weights in our deep architecture on external unlabeled datasets (like Wikipedia), than jointly train on all discourse parsing subtasks on the training set, use a separate validation set to optimize hyper-parameters, and estimate its performance on the test set. For evaluation purposes standard evaluation measures for subtasks based on F_1-scores will be used.

6 Conclusion

To increase the generality of our unified end-to-end approach for text-level discourse parsing, we will try to depend as little as possible on background knowledge in the form of hand-engineered features for a specific language, domain, or task. By incorporating various improvements in automatic learning of features and representations we hope to reach or even surpass performance of current state-of-the-art methods on annotated English corpora.

References

Jason Baldridge and Alex Lascarides. 2005. Probabilistic head-driven parsing for discourse structure. In *Proc. 9th Conf. Comput. Nat. Lang. Learn.*, pages 96–103. Association for Computational Linguistics.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. *Curr. New Dir. Discourse Dialogue*, 22:85–112.

Danqi Chen and Christopher D Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, pages 740–750.

Ronan Collobert and Jason Weston. 2008. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proc. 25th Int. Conf. Mach. Learn.*, volume 20, pages 160–167.

Ronan Collobert. 2011. Deep Learning for Efficient Discriminative Parsing. *Int. Conf. Artif. Intell. Stat.*, 15:224–232.

David A. DuVerle and Helmut Prendinger. 2009. A novel discourse parser based on support vector machine classification. In *Proc. Jt. Conf. 47th Annu. Meet. ACL 4th Int. Jt. Conf. Nat. Lang. Process. AFNLP*, pages 665–673. Association for Computational Linguistics.

Jeffrey L. Elman. 1990. Finding structure in time* 1. *Cogn. Sci.*, 14(1990):179–211.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level Discourse Parsing with Rich Linguistic Features. In *Proc. 50th Annu. Meet. Assoc. Comput. Linguist.*, pages 60–68. Association for Computational Linguistics.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence. In *Proc. 25th Int. Conf. Comput. Linguist.*

Seeger Fisher and Brian Roark. 2007. The utility of parse-derived features for automatic discourse segmentation. In *Proc. 45th Annu. Meet. Assoc. Comput. Linguist.*, volume 45, pages 488–495.

Sucheta Ghosh, Giuseppe Riccardi, and Richard Johansson. 2012. Global features for shallow discourse parsing. In *Annu. Meet. Spec. Interes. Gr. Discourse Dialogue*, pages 150–159.

Christoph Goller and Andreas Küchler. 1996. Learning Task-Dependent Distributed Representations by Backpropagation Through Structure. In *IEEE Int. Conf. Neural Networks*, pages 347–352.

Alex Graves, Greg Wayne, and Ivo Denihelka. 2014. Neural Turing Machines. *arXiv Prepr. arXiv410.5401*, pages 1–26.

Weiwei Guo and Mona Diab. 2012. Modeling Sentences in the Latent Space. In *Proc. 50th Annu. Meet. Assoc. Comput. Linguist.*, pages 864–872. Association for Computational Linguistics.

Michiel Hermans and Benjamin Schrauwen. 2013. Training and Analyzing Deep Recurrent Neural Networks. In *Adv. Neural Inf. Process. Syst.*, volume 26, pages 190–198.

Hugo Hernault, Helmut Prendinger, David A. DuVerle, and Mitsuru Ishizuka. 2010. HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue and Discourse*, 1(3):1–33.

Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.*, 18:1527–1554.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation Learning for Text-level Discourse Parsing. In *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist.*, pages 13–24.

Yangfeng Ji and Jacob Eisenstein. 2015. One Vector is Not Enough: Entity-Augmented Distributed Semantics for Discourse Relations. *Trans. Assoc. Comput. Linguist.*

Shafiq Joty and Alessandro Moschitti. 2014. Discriminative Reranking of Discourse Parses Using Tree Kernels. In *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process.*, pages 2049–2060.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proc. 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn.*, pages 904–915. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. 2013. Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis. In *Proc. 51st Annu. Meet. Assoc. Comput. Linguist.*, pages 486–496.

Fang Kong, Hwee Tou, and Ng Guodong. 2014. A Constituent-Based Approach to Argument Labeling with Joint Inference in Discourse Parsing. In *Conf. Empir. Methods Nat. Lang. Process.*, pages 68–77.

Jiwei Li, Dan Jurafsky, and Eduard Hovy. 2015. When Are Tree Structures Necessary for Deep Learning of Representations? *Arxiv*.

Junyi Jessy Li. 2014. Reducing Sparsity Improves the Recognition of Implicit Discourse Relations. In *Proc. SIGDIAL 2014 Conf.*, number June, pages 199–207.

Ziheng Lin, Min-yen Kan, and Hwee Tou Ng. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proc. 2009 Conf. Empir. Methods Nat. Lang. Process.*, pages 343–351. Association for Computational Linguistics.

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.

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

Daniel Marcu. 2000. The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Comput. Linguist.*, 26(38):395–448.

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.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Exploiting Scope for Shallow Discourse Parsing. In *Int. Conf. Lang. Resour. Eval.*, pages 2076–2083.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45(11):2673–2681.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic Compositionality through Recursive Matrix-Vector Spaces. In *Proc. 2012 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn.*, pages 1201–1211. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. Conf. Empir. Methods Nat. Lang. Process.*, pages 1631–1642.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. *Proc. 2003 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, 1:228–235.

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proc. Hum. Lang. Technol. 2009 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist.*, pages 566–574. Association for Computational Linguistics.

Mihai Surdeanu, Thomas Hicks, and Marco A. Valenzuela-Escarcega. 2015. Two Practical Rhetorical Structure Theory Parsers. In *Proc. North Am. Chapter Assoc. Comput. Linguist.*

Alexander Waibel, Toshiyuki Hanazawa, Geoffrey E. Hinton, Kiyohiro Shikano, and Kevin J. Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust.*, 37(3):328–339.