Spectral Semi-Supervised Discourse Relation Classification

Robert Fisher
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
rwfisher@cs.cmu.edu

Reid Simmons
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
reids@cs.cmu.edu

Abstract

Discourse parsing is the process of discovering the latent relational structure of a long form piece of text and remains a significant open challenge. One of the most difficult tasks in discourse parsing is the classification of implicit discourse relations. Most state-of-the-art systems do not leverage the great volume of unlabeled text available on the web–they rely instead on human annotated training data. By incorporating a mixture of labeled and unlabeled data, we are able to improve relation classification accuracy, reduce the need for annotated data, while still retaining the capacity to use labeled data to ensure that specific desired relations are learned. We achieve this using a latent variable model that is trained in a reduced dimensionality subspace using spectral methods. Our approach achieves an F_1 score of 0.485 on the implicit relation labeling task for the Penn Discourse Treebank.

1 Introduction

Discourse parsing is a fundamental task in natural language processing that entails the discovery of the latent relational structure in a multi-sentence piece of text. Unlike semantic and syntactic parsing, which are used for single sentence parsing, discourse parsing is used to discover inter-sentential relations in longer pieces of text. Without discourse, parsing methods can only be used to understand documents as sequences of unrelated sentences.

Unfortunately, manual annotation of discourse structure in text is costly and time consuming. Multiple annotators are required for each relation to estimate inter-annotator agreement. The Penn Discourse Treebank (PDTB) (Prasad et al., 2008). is one of the largest annotated discourse parsing datasets, with 16,224 implicit relations. However, this pales in comparison to unlabeled datasets that can include millions of sentences of text. By augmenting a labeled dataset with unlabeled data, we can use a bootstrapping framework to improve predictive accuracy, and reduce the need for labeled data–which could make it much easier to port discourse parsing algorithms to new domains. On the other hand, a fully unsupervised parser may not be desirable because in many applications specific discourse relations must be identified, which would be difficult to achieve without the use of labeled examples.

There has recently been growing interest in a breed of algorithms based on spectral decomposition, which are well suited to training with unlabeled data. Spectral algorithms utilize matrix factorization algorithms such as Singular Value Decomposition (SVD) and rank factorization to discover *low-rank decompositions* of matrices or tensors of empirical moments. In many models, these decompositions allow us to identify the subspace spanned by a group of parameter vectors or the actual parameter vectors themselves. For tasks where they can be applied, spectral methods provide statistically consistent results that avoid local maxima. Also, spectral algorithms tend to be much faster—sometimes orders of magnitude faster—than competing approaches, which makes them ideal for tackling large datasets. These methods can be viewed as inferring something about the latent structure of a domain—for example, in a hidden Markov model, the number of latent states and the sparsity pattern of the transition matrix are forms of latent structure, and spectral methods can recover both in the limit.

This paper presents a semi-supervised spectral model for a sequential relation labeling task for discourse parsing. Besides the theoretically desirable properties mentioned above, we also demon-

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 89–93,
Beijing, China, July 26-31, 2015.

strate the practical advantages of the model with an empirical evaluation on the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) dataset, which yields an F_1 score of 0.485. This accuracy shows a 7-9 percentage point improvement over approaches that do not utilize unlabeled training data.

2 Related Work

There has been quite a bit of work concerning fully supervised relation classification with the PDTB (Lin et al., 2014; Feng and Hirst, 2012; Webber et al., 2012). Semi-supervised relation classification is much less common however. One recent example of an attempt to leverage unlabeled data appears in (Hernault et al., 2011), which showed that moderate classification accuracy can be achieved with very small labeled datasets. However, this approach is not competitive with fully supervised classifiers when more training data is available. Recently there has also been some work to use Conditional Random Fields (CRFs) to represent the global properties of a parse sequence (Joty et al., 2013; Feng and Hirst, 2014), though this work has focused on the RST-DT corpus, rather than the PDTB.

In addition to requiring a fully supervised training set, most existing discourse parsers use non-spectral optimization that is often slow and inexact. However, there has been some work in other parsing tasks to employ spectral methods in both supervised and semi-supervised settings (Parikh et al., 2014; Cohen et al., 2014). Spectral methods have also been applied very successfully in many non-linguistic domains (Hsu et al., 2012; Boots and Gordon, 2010; Fisher et al., 2014).

3 Problem Definition and Dataset

This section defines the discourse parsing problem and discusses the characteristics of the PDTB. The PDTB consists of annotated articles from the Wall Street Journal and is used in our empirical evaluations. This is combined with the New York Times Annotated Corpus (Sandhaus, 2008), which includes 1.8 million New York Times articles printed between 1987 and 2007.

Discourse parsing can be reduced to three separate tasks. First, the text must be decomposed into *elementary discourse units* (EDUs), which may or may not coincide with sentence boundaries. The EDUs are often independent clauses that may be connected with conjunctions. After the text has been partitioned into EDUs, the discourse structure must be identified. This requires us to identify all pairs of EDUs that will be connected with *some* discourse relation. These relational links induce the skeletal structure of the discourse parse tree. Finally, each connection identified in the previous step must be labeled using a known set of relations. Examples of these discourse relations include concession, causal, and instantiation relations. In the PDTB, only adjacent discourse units are connected with a discourse relation, so with this dataset we are considering parse sequences rather than parse trees.

In this work, we focus on the relation labeling task, as fairly simple methods perform quite well at the other two tasks (Webber et al., 2012). We use the ground truth parse structures provided by the PDTB dataset, so as to isolate the error introduced by relation labeling in our results, but in practice a greedy structure learning algorithm can be used if the parse structures are not known *a priori*.

Some of the relations in the dataset are induced by specific connective words in the text. For example, a contrast relation may be explicitly revealed by the conjunction *but*. Simple classifiers using only the text of the discourse connective with POS tags can find explicit relations with high accuracy (Lin et al., 2014). The following sentence shows an example of a more difficult implicit relation. In this sentence, two EDUs are connected with an explanatory relation, shown in bold, although the connective word does not occur in the text.

> "But a few funds have taken other defensive steps. Some have raised their cash positions to record levels. **[BECAUSE]** High cash positions help buffer a fund when the market falls."

We focus on the more difficult implicit relations that are not induced by coordinating connectives in the text. The implicit relations have been shown to require more sophisticated feature sets including syntactic and linguistic information (Lin et al., 2009). The PDTB dataset includes 16,053 examples of implicit relations.

A full list of the PDTB relations is available in (Prasad et al., 2008). The relations are organized hierarchically into top level, types, and subtypes. Our experiments focus on learning only up

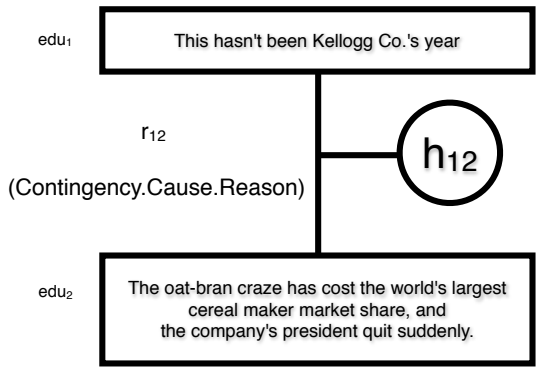

Figure 1: An example of the latent variable discourse parsing model taken from the Penn Discourse Treebank Dataset. The relation here is an example of a cause attribution relation.

to level 2, as the level 3 (sub-type) relations are too specific and show only 80% inter-annotator agreement. There are 16 level 2 relations in the PDTB, but the 5 least common relations only appear a handful of times in the dataset and are omitted from our tests, yielding 11 possible classes.

4 Approach

We incorporate unlabeled data into our spectral discourse parsing model using a bootstrapping framework. The model is trained over several iterations, and the most useful unlabeled sequences are added as labeled training data after each iteration. Our method also utilizes Markovian latent states to compactly capture global information about a parse sequence, with one latent variable for each relation in the discourse parsing sequence. Most discourse parsing frameworks will label relations independently of the rest of the accompanying parse sequence, but this model allows for information about the global structure of the discourse parse to be used when labeling a relation. A graphical representation of one link in the parsing model is shown in Figure 1.

Specifically, each potential relation r_{ij} between elementary discourse units e_i and e_j is accompanied by a corresponding latent variable as h_{ij}. According to the model assumptions, the following equality holds:

$$P(r_{ij} = r | r_{1,2}, r_{2,3} ... r_{n+1,n}) = P(r_{ij} = r | h_{ij})$$

To maintain notational consistency with other latent variable models, we will denote these relation variables as $x_1 ... x_n$, keeping in mind that there is one possible relation for each adjacent pair of elementary discourse units.

For the Penn Discourse Treebank Dataset, the discourse parses behave like sequence of random variables representing the relations, which allows us to use an HMM-like latent variable model based on the framework presented in (Hsu et al., 2012). If the discourse parses were instead trees, such as those seen in Rhetorical Structure Theory (RST) datasets, we can modify the standard model to include separate parameters for left and right children, as demonstrated in (Dhillon et al., 2012).

4.1 Spectral Learning

This section briefly describes the process of learning a spectral HMM. Much more detail about the process is available in (Hsu et al., 2012). Learning in this model will occur in a subspace of dimensionality m, but system dynamics will be the same if m is not less than the rank of the observation matrix. If our original feature space has dimensionality n, we define a transformation matrix $U \in \mathbb{R}^{n \times m}$, which can be computed using Singular Value Decomposition. Given the matrix U, coupled with the empirical unigram (P_1), bigram ($P_{2,1}$), and trigram matrices ($P_{3,x,1}$), we are able to estimate the subspace initial state distribution ($\hat{\pi}_U$) and observable operator ($\hat{A}_U$) using the following equalities (wherein the Moore-Penrose pseudo-inverse of matrix X is denoted by X^+):

$$\hat{\pi}_U = U^T P_1$$
$$\hat{A}_U = U^T P_{3,x,1} (U^T P_{2,1})^+ \; \forall x$$

For our original feature space, we use the rich linguistic discourse parsing features defined in (Feng and Hirst, 2014), which includes syntactic and linguistic features taken from dependency parsing, POS tagging, and semantic similarity measures. We augment this feature space with a vector space representation of semantics. A term-document co-occurrence matrix is computed using all of Wikipedia and Latent Dirichlet Analysis was performed using this matrix. The top 200 concepts from the vector space representation for each pair of EDUs in the dataset are included in the feature space, with a concept regularization parameter of 0.01.

4.2 Semi-Supervised Training

To begin semi-supervised training, we perform a syntactic parse of the unlabeled data and ex-

tract EDU segments using the method described in (Feng and Hirst, 2014). The model is then trained using the labeled dataset, and the unlabeled relations are predicted using the Viterbi algorithm. The most informative sequences in the unlabeled training set are added to the labeled training set as labeled examples. To measure how informative a sequence of relations is, we use *density-weighted certainty sampling* (DCS). Specifically for a sequence of relations $r_1 ... r_n$ taken from a document, d, we use the following formula:

$$DCS(d) = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{p}(r_i)}{H(r_i)}$$

In this equation, $H(r_i)$ represented the entropy of the distribution of label predictions for the relation r_i generated by the current spectral model, which is a measure of the model's uncertainty for the label of the given relation. Density is denoted $\hat{p}(r_i)$, and this quantity measures the extent to which the text corresponding to this relation is representative of the labeled corpus. To compute this measure, we create a Kernel Density Estimate (KDE) over a 100 dimensional LDA vector space representation of all EDU's in the labeled corpus. We then compute the density of the KDE for the text associated with relation r_i, which gives us $\hat{p}(r_i)$. All sequences of relations in the unlabeled dataset are ranked according to their average density-weighted certainty score, and all sequences scoring above a parameter ψ are added to the training set. The model is then retrained, the unlabeled data re-scored, and the process is repeated for several iterations. In iteration i, the labeled data in the training set is weighted w_i^l, and the unlabeled data is weighted w_i^u, with the unlabeled data receiving higher weight in subsequent iterations. The KDE kernel bandwidth and the parameters ψ, w_i^l, w_i^u, and the number of hidden states are chosen in experiments using 10-fold cross validation on the labeled training set, coupled with a subset of the unlabeled data.

5 Results

Figure 2 shows the F_1 scores of the model using various sizes of labeled training sets. In all cases, the entirety of the unlabeled data is made available, and 7 rounds of bootstrapping is conducted. Sections 2-22 of the PDTB are used for training, with section 23 being withheld for testing, as recommended by the dataset guidelines (Prasad et al., 2008). The results are compared against those reported in (Lin et al., 2014), as well as a simple baseline classifier that labels all relations with the most common class, $EntRel$. Compared to the semi-supervised method described in (Hernault et al., 2011), we show significant gains in accuracy at various sizes of dataset, although the unlabeled dataset used in our experiments is much larger.

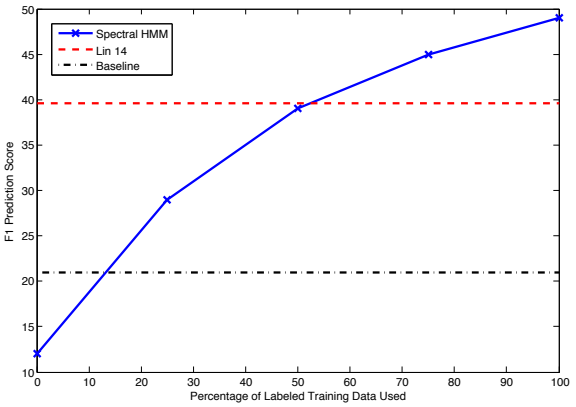

Figure 2: Empirical results for labeling of implicit relations.

When the spectral HMM is trained using only the labeled dataset, with no unlabeled data, it produces an F_1 score of 41.1%, which is comparable to the results reported in (Lin et al., 2014). By comparison, the semi-supervised classifier is able to obtain similar accuracy when using approximately 50% of the labeled training data. When given access to the full labeled dataset, we see an improvement in the F_1 score of 7-9 percentage points. Recent work has shown promising results using CRFs for discourse parsing (Joty et al., 2013; Feng and Hirst, 2014), but the results reported in this work were taken from the RST-DT corpus and are not directly comparable. However, supervised CRFs and HMMs show similar accuracy in other language tasks (Ponomareva et al., 2007; Awasthi et al., 2006).

6 Conclusions

In this work, we have shown that we are able to outperform fully-supervised relation classifiers by augmenting the training data with unlabeled text. The spectral optimization used in this approach makes computation tractable even when using over one million documents. In future work, we would like to further improve the performance of this method when very small labeled training

sets are available, which would allow discourse analysis to be applied in many new and interesting domains.

Acknowledgements

We give thanks to Carolyn Penstein Rosé and Geoff Gordon for their helpful discussions and suggestions. We also gratefully acknowledge the National Science Foundation for their support through EAGER grant number IIS1450543. This material is also based upon work supported by the Quality of Life Technology Center and the National Science Foundation under Cooperative Agreement EEC-0540865.

References

Pranjal Awasthi, Delip Rao, and Balaraman Ravindran. 2006. Part of speech tagging and chunking with hmm and crf. *Proceedings of NLP Association of India (NLPAI) Machine Learning Contest 2006*.

Byron Boots and Geoffrey J Gordon. 2010. Predictive state temporal difference learning. *arXiv preprint arXiv:1011.0041*.

Shay B Cohen, Karl Stratos, Michael Collins, Dean P Foster, and Lyle Ungar. 2014. Spectral learning of latent-variable pcfgs: Algorithms and sample complexity. *The Journal of Machine Learning Research*, 15(1):2399–2449.

Paramveer S Dhillon, Jordan Rodu, Michael Collins, Dean P Foster, and Lyle H Ungar. 2012. Spectral dependency parsing with latent variables. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 205–213. 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*, pages 60–68. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of The 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, USA, June*.

Robert Fisher, Reid Simmons, Cheng-Shiu Chung, Rory Cooper, Garrett Grindle, Annmarie Kelleher, Hsinyi Liu, and Yu Kuang Wu. 2014. Spectral machine learning for predicting power wheelchair exercise compliance. In *Foundations of Intelligent Systems*, pages 174–183. Springer.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2011. Semi-supervised discourse relation classification with structural learning. In *Computational Linguistics and Intelligent Text Processing*, pages 340–352. Springer.

Daniel Hsu, Sham M Kakade, and Tong Zhang. 2012. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480.

Shafiq R Joty, Giuseppe Carenini, Raymond T Ng, and Yashar Mehdad. 2013. Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In *ACL (1)*, pages 486–496.

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*, 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. *Natural Language Engineering*, pages 1–34.

Ankur Parikh, Shay B Cohen, and Eric Xing. 2014. Spectral unsupervised parsing with additive tree metrics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Long Papers*. Association for Computational Linguistics.

Natalia Ponomareva, Paolo Rosso, Ferrán Pla, and Antonio Molina. 2007. Conditional random fields vs. hidden markov models in a biomedical named entity recognition task. In *Proc. of Int. Conf. Recent Advances in Natural Language Processing, RANLP*, pages 479–483.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Evan Sandhaus. 2008. The new york times annotated corpus ldc2008t19. *Linguistic Data Consortium*.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.