# Coherence Modeling Improves Implicit Discourse Relation Recognition

**Noriki Nishida** and **Hideki Nakayama**
Graduate School of Information Science and Technology
The University of Tokyo
{nishida,nakayama}@nlab.ci.i.u-tokyo.ac.jp

## Abstract

The research described in this paper examines how to learn linguistic knowledge associated with discourse relations from unlabeled corpora. We introduce an unsupervised learning method on text coherence that could produce numerical representations that improve implicit discourse relation recognition in a semi-supervised manner. We also empirically examine two variants of coherence modeling: *order-oriented* and *topic-oriented* negative sampling, showing that, of the two, topic-oriented negative sampling tends to be more effective.

## 1 Introduction

Shallow discourse parsing aims to automatically identify discourse relations (e.g., comparisons) between adjacent sentences. When connectives such as *however* explicitly appear, discourse relations are relatively easy to classify, as connectives provide strong cues (Pitler et al., 2008). In contrast, it remains challenging to identify discourse relations across sentences that have no connectives.

One reason for this inferior performance is a shortage of labeled instances, despite the diversity of natural language discourses. Collecting annotations about implicit relations is highly expensive because it requires linguistic expertise. [1] A variety of semi-supervised or unsupervised methods have been explored to alleviate this issue. Marcu and Echihabi (2002) proposed generating synthetic instances by removing connectives from sentence pairs. This idea has been extended in many works and remains a core approach in the field (Zhou et al., 2010; Patterson and Kehler, 2013; Lan et al., 2013; Rutherford and Xue, 2015; Ji et al., 2015; Liu et al., 2016; Braud and Denis, 2016; Lan et al., 2017; Wu et al., 2017). However, these methods rely on automatically detecting connectives in unlabeled corpora beforehand, which makes it almost impossible to utilize parts of unlabeled corpora in which no connectives appear.[2] In addition, as Sporleder and Lascarides (2008) discovered, it is difficult to obtain a generalized model by training on synthetic data due to domain shifts. Though several semi-supervised methods do not depend on detecting connectives (Hernault et al., 2010, 2011; Braud and Denis, 2015), these methods are restricted to manually selected features, linear models, or word-level knowledge transfer.

In this paper, our research question is how to exploit unlabeled corpora without explicitly detecting connectives to learn linguistic knowledge associated with implicit discourse relations.

Our core hypothesis is that unsupervised learning about text coherence could produce numerical representations related to discourse relations. Sentences that compose a coherent document should be connected with syntactic or semantic relations (Hobbs, 1985; Grosz et al., 1995). In particular, we expect that there should be latent relations among local sentences. In this study, we hypothesize that parameters learned through coherence modeling could contain useful information for identifying (implicit) discourse relations. To verify this hypothesis, we develop a semi-supervised system whose parameters are first optimized for coherence modeling and then transferred to implicit discourse relation recognition. We also empirically examine two variants of coherence mod-

[1] The Penn Discourse Treebank (PDTB) 2.0 corpus (Prasad et al., 2008), which is the current largest corpus for discourse relation recognition, contains only about 16K annotated instances in total.

[2] For example, nearly half of the sentences in the British National Corpus hold implicit discourse relations and do not contain connectives (Sporleder and Lascarides, 2008).



*Proceedings of the SIGDIAL 2018 Conference*, pages 344–349,
Melbourne, Australia, 12-14 July 2018. 

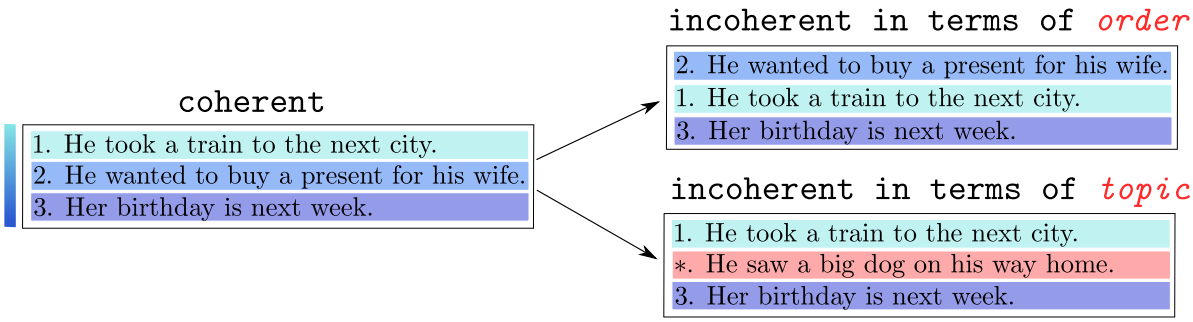


Figure 1: An example of order-oriented and topic-oriented negative sampling in coherence modeling.

eling: (1) *order-oriented* negative sampling and (2) *topic-oriented* negative sampling. An example is shown in Figure 1.

Our experimental results demonstrate that coherence modeling improves Macro $F_1$ on implicit discourse relation recognition by about 3 points on first-level relation *classes* and by about 5 points on second-level relation *types*. Coherence modeling is particularly effective for relation categories with few labeled instances, such as temporal relations. In addition, we find that topic-oriented negative sampling tends to be more effective than the order-oriented counterpart, especially on first-level relation classes.

## 2 Coherence Modeling

In this study, we adopt the sliding-window approach of Li and Hovy (2014) to form a conditional probability that a document is coherent. That is, we define the probability that a given document $X$ is coherent as a product of probabilities at all possible local windows, i.e.,

$$P(\text{coherent}|X,\theta) = \prod_{x \in X} P(\text{coherent}|x,\theta), \quad (1)$$

where $P(\text{coherent}|x,\theta)$ denotes the conditional probability that the local clique $x$ is coherent and $\theta$ denotes parameters. Clique $x$ is a tuple of a central sentence and its left and right sentences, $(s_-, s, s_+)$. Though larger window sizes may allow the model to learn linguistic properties and inter-sentence dependencies over broader contexts, it increases computational complexity during training and suffers from data sparsity problem.

We automatically build a dataset $\mathcal{D} = \mathcal{P} \cup \mathcal{N}$ for coherence modeling from an unlabeled corpus. Here, $\mathcal{P}$ and $\mathcal{N}$ denote sets of positive and negative instances, respectively. Given a source corpus $\mathcal{C}$ of $|\mathcal{C}|$ sentences $s_1, s_2, \ldots, s_{|\mathcal{C}|}$, we collect positive instances as follows:

$$\mathcal{P} = \{(s_{i-1}, s_i, s_{i+1}) \mid i = 2, \ldots, |\mathcal{C}| - 1\}. \quad (2)$$

Text coherence can be corrupted by two aspects, which correspond to how to build negative set $\mathcal{N}$.

The first variant is ***order**-oriented negative sampling*, i.e.,

$$\mathcal{N} = \{x' \mid x' \in \phi(x) \wedge x \in \mathcal{P}\} \quad (3)$$

where $\phi(x)$ denotes the set of possible permutations of $x$, excluding $x$ itself.

The second variant is ***topic**-oriented negative sampling*, i.e.,

$$\mathcal{N} = \{(s_-, s', s_+) \mid s' \in \mathcal{C} \wedge (s_-, s, s_+) \in \mathcal{P}\} \quad (4)$$

where $s'$ denotes a sentence randomly sampled from a uniform distribution over the entire corpus $\mathcal{C}$. We call this method *topic-oriented* because topic consistency shared across a clique $(s_-, s, s_+)$ is expected to be corrupted by replacing $s$ with $s'$.

## 3 Model Architecture

We develop a simple semi-supervised model with neural networks. An overall view is shown in Figure 2. Our model mainly consists of three components: sentence encoder $E$, coherence classifier $F_c$, and implicit discourse relation classifier $F_r$. The parameters of $E$ are shared across the two tasks: coherence modeling and implicit discourse relation recognition. In contrast, $F_c$ and $F_r$ are optimized separately. Though it is possible to develop more complex architectures (such as with word-level matching (Chen et al., 2016), a soft-attention mechanism (Liu and Li, 2016; Rönnqvist et al., 2017), or highway connections (Qin et al.,

| | 1st-Level Relation *Classes* | | 2nd-Level Relation *Types* | | Coherence |
|---|---|---|---|---|---|
| | Acc. (%) | Macro $F_1$ (%) | Acc. (%) | Macro $F_1$ (%) | Acc. (%) |
| *IRel* only | 51.49 | 42.29 | 37.49 | 24.81 | N/A |
| *IRel* + *O-Coh* (Small) | 52.16 | 41.39 | 37.77 | 25.46 | 57.96 |
| *IRel* + *O-Coh* (Large) | 52.29 | 42.48 | 41.29 | **30.70** | **64.24** |
| *IRel* + *T-Coh* (Small) | 51.70 | 40.84 | 37.91 | 25.35 | 83.04 |
| *IRel* + *T-Coh* (Large) | **53.54** | **45.03** | **41.39** | 29.67 | **91.53** |

Table 1: The results of implicit discourse relation recognition (multi-class classification) and coherence modeling (binary classification). *IRel* and *O/T-Coh* denote that the model is trained on implicit discourse relation recognition and order/topic-oriented coherence modeling respectively. "Small" and "large" correspond to the relative size of the used unlabeled corpus: 37K (WSJ) and 22M (BLLIP) positive instances, respectively.

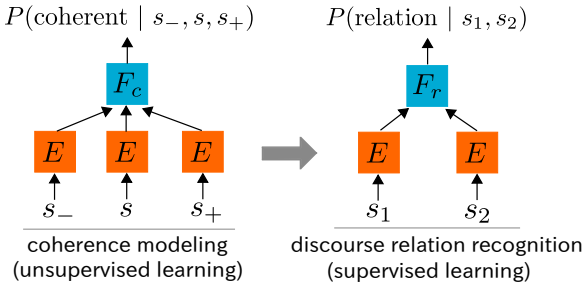


Figure 2: The semi-supervised system we developed. The model consists of sentence encoder $E$, coherence classifier $F_c$, and implicit discourse relation classifier $F_r$.

2016)), such architectures are outside the scope of this study, since the effectiveness of incorporating coherence-based knowledge would be broadly orthogonal to the model's complexity.

### 3.1 Sentence Encoder

Sentence encoder $E$ transforms a symbol sequence (i.e., a sentence) into a continuous vector. First, a bidirectional LSTM (BiLSTM) is applied to a given sentence of $n$ tokens $w_1, \ldots, w_n$, i.e.,

$$\overrightarrow{\boldsymbol{h}}_i = \text{FwdLSTM}(\overrightarrow{\boldsymbol{h}}_{i-1}, w_i) \in \mathbb{R}^D, \quad (5)$$

$$\overleftarrow{\boldsymbol{h}}_i = \text{BwdLSTM}(\overleftarrow{\boldsymbol{h}}_{i+1}, w_i) \in \mathbb{R}^D \quad (6)$$

where FwdLSTM and BwdLSTM denote forward and backward LSTMs, respectively. We initialize the hidden states to zero vectors, i.e., $\overrightarrow{\boldsymbol{h}}_0 = \overleftarrow{\boldsymbol{h}}_{n+1} = \mathbf{0}$. In our preliminary experiments, we tested conventional pooling functions (e.g., summation, average, or maximum pooling); we found that the following concatenation tends to yield higher performances:

$$\boldsymbol{h} = \left(\overrightarrow{\boldsymbol{h}}_L^\top, \overleftarrow{\boldsymbol{h}}_1^\top\right)^\top \in \mathbb{R}^{2D}. \quad (7)$$

We use Eq. 7 as the aggregation function throughout our experiments.

### 3.2 Classifiers

We develop two multi-layer perceptrons (MLPs) with ReLU nonlinearities followed by softmax normalization each for $F_c$ and $F_r$. The MLP inputs are the concatenation of sentence vectors. Thus, the dimensionalities of the input layers are $2D \times 3$ and $2D \times 2$ respectively. The MLPs consist of `input`, `hidden`, and `output` layers.

## 4 Experiments

### 4.1 Preparation

We used the Penn Discourse Treebank (PDTB) 2.0 corpus (Prasad et al., 2008) as a dataset for implicit discourse relation recognition. We followed the standard section partition, which is to use Sections 2–20 for training, Sections 0-1 for development, and Sections 21–22 for testing. We evaluate multi-class classifications with first-level relation *classes* (four classes) and second-level relation *types* (11 classes).

We used the Wall Street Journal (WSJ) articles (Marcus et al., 1993)[3] or the BLLIP North American News Text (Complete) (McClosky et al., 2008)[4] to build a coherence modeling dataset, resulting in about 48K (WSJ) or 23M (BLLIP) positive instances. We inserted a special symbol "⟨ARTICLE_BOUNDARY⟩" to each

[3]We used the raw texts in LDC99T42 Treebank-3: https://catalog.ldc.upenn.edu/LDC99T42

[4]https://catalog.ldc.upenn.edu/LDC2008T13

| | Acc. (%) | Macro $F_1$ (%) |
|---|---|---|
| Rutherford and Xue (2015) | 57.10 | 40.50 |
| Liu et al. (2016) | 57.27 | 44.98 |
| Braud and Denis (2016)[5] | 52.81 | 42.27 |
| Wu et al. (2017) | **58.85** | 44.84 |
| *IRel* only | 51.49 | 42.29 |
| *IRel* only* | 52.72 | 42.61 |
| *IRel* + *T-Coh* (Large) | 53.54 | **45.03** |
| *IRel* + *T-Coh* (Large)* | 56.60 | **46.90** |

Table 2: Comparison with previous works that exploit unlabeled corpora on first-level relation *classes*. An asterisk indicates that word embeddings are fine-tuned (which slightly decreases performance on second-level relation *types* due to overfitting).

| | Exp. | Cont. | Comp. | Temp. |
|---|---|---|---|---|
| # of training data | 6,673 | 3,235 | 1,855 | 582 |
| *IRel* only | 66.40 | 53.49 | 39.48 | 32.31 |
| *IRel* + *T-Coh* | **67.48** | **54.94** | **40.41** | **35.60** |

Table 3: Results on one-vs.-others binary classification in implicit discourse relation recognition. The evaluation metric is Macro $F_1$ (%). We evaluate on the first-level relation *classes*: `Expansion`, `Contingency`, `Comparison`, and `Temporal`.

article boundary. For the WSJ corpus, we split the sections into training/development/test sets in the same way with the implicit relation recognition. For the BLLIP corpus, we randomly sampled 10,000 articles each for the development and test sets. Negative instances are generated following the procedure described in Section 2. Note that this procedure requires neither human annotation nor special connective detection.

We set the dimensionalities of the word embeddings, hidden states of the BiLSTM, and hidden layers of the MLPs to 100, 200, and 100, respectively. GloVe (Pennington et al., 2014) was used to produce pre-trained word embeddings on the BLLIP corpus. To avoid overfitting, we fixed the word embeddings during training in both coherence modeling and implicit relation recognition. Dropout (ratio 0.2) was applied to word embeddings and MLPs's layers. At every iteration during training in both tasks, we configured class-balanced batches by resampling.

[5]The values are taken from Wu et al. (2017).

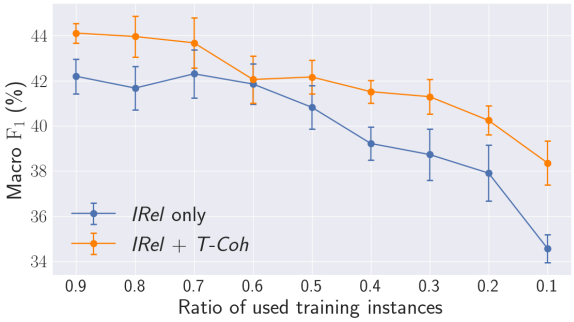


Figure 3: Results on implicit discourse relation recognition (first-level *classes*), with different numbers of training instances. The error bars show one standard deviation over 10 trials.

### 4.2 Results

To verify whether unsupervised learning on coherence modeling could improve implicit discourse relation recognition, we compared the semi-supervised model (i.e., implicit discourse relation recognition (***IRel***) + coherence modeling with order/topic-oriented negative sampling (***O/T-Coh***)) with the baseline model (i.e., *IRel* only). The evaluation metrics are accuracy (%) and Macro $F_1$ (%). We report the mean scores over 10 trials. Table 1 shows that coherence modeling improves Macro $F_1$ by about 3 points in first-level relation *classes* and by about 5 points in second-level relation *types*. Coherence modeling also outperforms the baseline in accuracy. We observed that the higher the coherence modeling performance (see Small vs. Large), the higher the implicit relation recognition score. These results support our claim that coherence modeling could learn linguistic knowledge that is useful for identifying discourse relations.

We also found that topic-oriented negative sampling tends to outperform its order-oriented counterpart, especially on first-level relation *classes*. We suspect that this is because order-oriented coherence modeling is more fine-grained and challenging than topic-oriented identification, resulting in poor generalization. For example, there could be order-invariant cliques that still hold coherence relations after random shuffling, whereas topic-invariant cliques hardly exist. Indeed, training on order-oriented negative sampling converged to lower scores than that of topic-oriented negative sampling (see coherence accuracy).

Next, for reference, we compared our system with previous work that exploits unlabeled cor-

pora. As shown in Table 2, we found our model to outperform previous systems in Macro $F_1$. In this task, Macro $F_1$ is more important than accuracy because the class balance in the test set is highly skewed. Note that these previous models rely on previously detected connectives in the unlabeled corpus, whereas our system is free from such detection procedures.

To assess the effectiveness of coherence modeling on different relation classes, we trained and evaluated the models on one-vs-others binary classification. That is, we treated each of the first-level relation *classes* (4 classes) as the positive class and others as the negative class. Table 3 shows that coherence modeling is effective, especially for the `Temporal` relation which has relatively fewer labeled instances than others, indicating that coherence modeling could compensate for the shortage of labeled data.

We also performed an ablation study to discover the performance contribution from coherence modeling by changing the number of training instances used in implicit relation recognition. Here, we assume that in real-world situations, we do not have sufficient labeled data. We downsampled from the original training set and maintained the balance of classes as much as possible. As shown in Figure 3, coherence modeling robustly yields improvements, even if we reduced the labeled instances to 10%.

## 5 Conclusion

In this paper, we showed that unsupervised learning on coherence modeling improves implicit discourse relation recognition in a semi-supervised manner. Our approach does not require detecting explicit connectives, which makes it possible to exploit entire unlabeled corpora. We empirically examined two variants of coherence modeling and show that topic-oriented negative sampling tends to be more effective than the order-oriented counterpart on first-level relation *classes*.

It still remains unclear whether the coherence-based knowledge is complemental to those by previous work. It is also interesting to qualitatively inspect the differences of learned properties between order-oriented and topic-oriented negative sampling. We will examine this line of research in future.

## Acknowledgments

The authors would like to thank anonymous reviewers for their constructive and helpful suggestions on this work. This work was supported by JSPS KAKENHI Grant Number 16H05872.

## References

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*.

Chloé Braud and Pascal Denis. 2016. Learning connective-based word representations for implicit discourse relation identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*.

Jifan Chen, Qi Zhang, Pengfei Liu, and Xuanjing Huang. 2016. Discourse relations detection via a mixed generative-discriminative framework. In *Proceedings of the 30th Conference on Artificial Intelligence (AAAI 2016)*.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistis*, 21(2):203–225.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2011. Semi-supervised discourse relation classification with structure learning. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)*.

Jerry R. Hobbs. 1985. On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information (CSLI), Stanford University.

Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. 2015. Closing the gap: domain adaptation from explicit to implicit discourse relations. In *Proceedings of the 53st Annual Meeting of the Association for Computational Linguistics (ACL 2015)*.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference of Empirical Methods in Natural Language Processing (EMNLP 2017)*.

Man Lan, Yu Xu, and Zhengyu Niu. 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.

Jiwei Li and Eduard Hovy. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: neural networks with multi-level attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the 30th Conference on Artificial Intelligence (AAAI 2016)*.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

David McClosky, Eugene Charniak, and Mark Johnson. 2008. Bllip north american news text, complete. *Linguistic Data Consortium*.

Gary Patterson and Andrew Kehler. 2013. Predicting the presence of discourse connectives. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2008)*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*.

Samuel Rönnqvist, Niko Schenk, and Christian Chiarcos. 2017. A recurrent neural model with attention for the recognition of chinese implicit discourse relations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

Attapol T. Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2015)*.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: an assessment. *Natural Language Engineering*, 14(03).

Changxing Wu, Xiaodong Shi, Yidong Chen, Jinsong Su, and Boli Wang. 2017. Improving implicit discourse relation recognition with discourse-specific word embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

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 (COLING 2010)*.