The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence

Vanessa Wei Feng[1], **Ziheng Lin**[2], **and Graeme Hirst**[1]
[1] Department of Computer Science
University of Toronto
{weifeng, gh}@cs.toronto.edu
[2] Singapore Press Holdings
linziheng@gmail.com

Abstract

Previous work by Lin et al. (2011) demonstrated the effectiveness of using discourse relations for evaluating text coherence. However, their work was based on discourse relations annotated in accordance with the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), which encodes only very shallow discourse structures; therefore, they cannot capture long-distance discourse dependencies. In this paper, we study the impact of deep discourse structures for the task of coherence evaluation, using two approaches: (1) We compare a model with features derived from discourse relations in the style of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), which annotate the full hierarchical discourse structure, against our re-implementation of Lin et al.'s model; (2) We compare a model encoded using only shallow RST-style discourse relations, against the one encoded using the complete set of RST-style discourse relations. With an evaluation on two tasks, we show that deep discourse structures are truly useful for better differentiation of text coherence, and in general, RST-style encoding is more powerful than PDTB-style encoding in these settings.

1 Introduction

In a well-written text, utterances are not simply presented in an arbitrary order; rather, they are presented in a logical and coherent form, so that the readers can easily interpret the meaning that the writer wishes to present. Therefore, coherence is one of the most essential aspects of text quality. Given its importance, the automatic evaluation of text coherence is one of the crucial components of many NLP applications.

A particularly popular model for the evaluation of text coherence is the entity-based local coherence model of Barzilay and Lapata (B&L) (2005; 2008), which extracts mentions of entities in the text, and models local coherence by the transitions, from one sentence to the next, in the grammatical role of each mention. Since the initial publication of this model, a number of extensions have been proposed, the majority of which are focused on enriching the original feature set. However, these enriched feature sets are usually application-specific, i.e., it requires a certain expertise and intuition to conceive good features.

In contrast, we seek insights of better feature encoding from a more general problem: discourse parsing (to be introduced in Section 2). Discourse parsing aims to identify the discourse relations held among various discourse units in the text. Therefore, one can expect that discourse parsing provides useful information to the evaluation of text coherence, because, essentially, the existence and the distribution of discourse relations are the basis of the coherence in a text.

In fact, there is already evidence showing that discourse relations can help better capture text coherence. Lin et al. (2011) use a PDTB-style discourse parser (to be introduced in Section 2.1) to identify discourse relations in the text, and they represent a text by entities and their associated discourse roles in each sentence. In their experiments, using discourse roles alone, their model performs very similar or even better than B&L's model. Combining their discourse role features with B&L's entity-based transition features further improves the performance.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 940–949, Dublin, Ireland, August 23-29 2014.

S_1: The dollar finished lower yesterday, after tracking another rollercoaster session on Wall Street.
S_2: [Concern about the volatile U.S. stock market had faded in recent sessions]$C_{2.1}$, [and traders appeared content to let the dollar languish in a narrow range until tomorrow, when the preliminary report on third-quarter U.S. gross national product is released.]$C_{2.2}$
S_3: But seesaw gyrations in the Dow Jones Industrial Average yesterday put Wall Street back in the spotlight and inspired market participants to bid the U.S. unit lower.

Three discourse relations are presented in the text above:

1. Implicit *EntRel* between S_1 as Arg1, and S_2 as Arg2.
2. Explicit *Conjunction* within S_2: $C_{2.1}$ as Arg1, $C_{2.2}$ as Arg2, with *and* as the connective.
3. Explicit *Contrast* between S_2 as Arg1 and S_3 as Arg2, with *but* as the connective.

Figure 1: An example text fragment composed of three sentences, and its PDTB-style discourse relations.

However, PDTB-style discourse relations encode only very shallow discourse structures, i.e., the relations are mostly local, e.g., within a single sentence or between two adjacent sentences. Therefore, in general, features derived from PDTB-style discourse relations cannot capture long discourse dependency, and thus the resulting model is still limited to being a local model. Nonetheless, long-distance discourse dependency could be quite useful for capturing text coherence from a global point of view.

Therefore, in this paper, we study the effect of deep hierarchical discourse structure in the evaluation of text coherence, by adopting two approaches to perform a direct comparison between models that incorporate deep hierarchical discourse structures and models with shallow structures. To evaluate our models, we conduct experiments on two datasets, each of which resembles a real sub-task in the evaluation of text coherence: **sentence ordering** and **essay scoring**. On both tasks, the model derived from deep discourse structures is shown to be more powerful than the model derived from shallow discourse structures. Moreover, for sentence ordering, combining our model with entity-based transition features achieves the best performance. However, for essay scoring, the combination is detrimental.

2 Discourse parsing

Discourse parsing is the problem of identifying the discourse structure within a text, by recognizing the specific type of its discourse relations, such as *Contrast*, *Explanation*, and *Causal* relations. Although discourse parsing is still relatively less well-studied, a number of theories have been proposed to capture different rhetorical characteristics or to serve different applications.

Currently, the two main directions in the study of discourse parsing are PDTB-style and RST-style parsing. These two directions are based on distinct theoretical frameworks, and each can be potentially useful for particular kinds of downstream applications. As will be discussed shortly, the major difference between PDTB- and RST-style discourse parsing is the notion of deep hierarchical discourse structure, which, according to our hypothesis, can be very useful for recognizing text coherence.

2.1 PDTB-style Discourse Parsing

The Penn Discourse Treebank (PDTB), developed by Prasad et al. (2008), is currently the largest discourse-annotated corpus, consisting of 2159 Wall Street Journal articles. The annotation in PDTB adopts the predicate-argument view of discourse relations, where a discourse connective (e.g., *because*) is treated as a predicate that takes two text spans as its arguments. The argument that the discourse connective structurally attaches to is called Arg2, and the other argument is called Arg1. In PDTB, relations are further categorized into *explicit* and *implicit* relations: a relation is explicit if there is an explicit discourse connective presented in the text; otherwise, it is implicit. PDTB relations focus more on *locality* and *adjacency*: explicit relations seldom connect text units beyond local context; for implicit relations,

S_1: [The dollar finished lower yesterday,]e_1 [after tracking another rollercoaster session on Wall Street.]e_2
S_2: [Concern about the volatile U.S. stock market had faded in recent sessions,]e_3 [and traders appeared content to let the dollar languish in a narrow range until tomorrow,]e_4 [when the preliminary report on third-quarter U.S. gross national product is released.]e_5
S_3: [But seesaw gyrations in the Dow Jones Industrial Average yesterday put Wall Street back in the spotlight]e_6 [and inspired market participants to bid the U.S. unit lower.]e_7

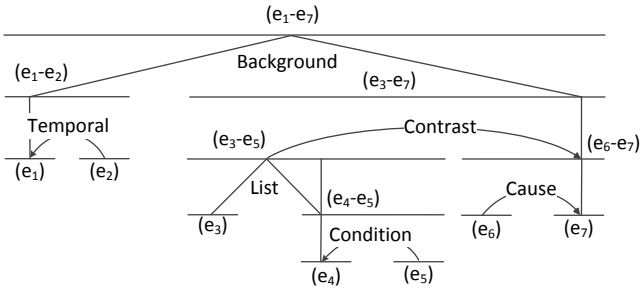

Figure 2: An example text fragment composed of seven EDUs, and its RST discourse tree representation.

only adjacent sentences within paragraphs are examined for the existence of implicit relations.

The PDTB-style discourse parsing is thus the type of framework in accordance with the PDTB, which extracts the discourse relations in a text, by identifying the presence of discourse connectives, the associated discourse arguments, and the specific types of the relations. An example text fragment is shown in Figure 1, consisting of three sentences, S_1, S_2, and S_3. A sentence may further contain clauses, e.g., $C_{2.1}$ and $C_{2.2}$ in S_2. The three PDTB-style discourse relations in this text are explained below the text.

2.2 RST-style Discourse Parsing

RST-style discourse parsing follows the theoretical framework of Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). In the framework of RST, a coherent text can be represented as a discourse tree whose leaves are non-overlapping text spans called *elementary discourse units* (EDUs); these are the minimal text units of discourse trees. Adjacent nodes can be related through particular discourse relations to form a discourse subtree, which can then be related to other adjacent nodes in the tree structure. RST-style discourse relations can be categorized into two types: mononuclear and multi-nuclear. In mononuclear relations, one of the text spans, the *nucleus*, is more salient than the other, the *satellite*, while in multi-nuclear relations, all text spans are equally important for interpretation.

Consider Figure 2, in which the same example as in Figure 1 is chunked into seven EDUs (e_1-e_7), segmented by square brackets. Its discourse tree representation is shown below in the figure, following the notational convention of RST. The two EDUs e_1 and e_2 are related by a mononuclear relation *Temporal*, where e_1 is the more salient span; e_4 and e_5 are related by *Condition*, with e_4 as the nucleus; and e_6 and e_7 are related by *Cause*, with e_7 as the nucleus. Then, the spans (e_3-e_5) and (e_6-e_7) are related by *Contrast* to form a higher-level discourse structure, and so on. Finally, a *Background* relation merges the span (e_1-e_2) and (e_3-e_7) on the top level of the tree.

As can be seen, thanks to the tree-structured representation of RST, compared to PDTB-style representation, we have a full hierarchy of discourse relations in the text: discourse relations exist not only in a local context, but also on higher text levels, such as between S_1 and the concatenation of S_2 and S_3.

3 Entity-based Local Coherence Model

The entity-based local coherence model was initially developed by Barzilay and Lapata (B&L) (2005; 2008). The fundamental assumption of this model is that a document makes repeated reference to elements of a set of entities that are central to its topic.

For a document d, an entity grid is constructed, in which the columns represent the entities referred

S_1: [**The dollar**]$_{\mathbf{S}}$ finished lower [**yesterday**]$_{\mathbf{X}}$, after tracking [**another rollercoaster session**]$_{\mathbf{O}}$ on [**Wall Street**]$_{\mathbf{X}}$.
S_2: [**Concern**]$_{\mathbf{S}}$ about [**the volatile U.S. stock market**]$_{\mathbf{X}}$ had faded in [**recent sessions**]$_{\mathbf{X}}$, and [**traders**]$_{\mathbf{S}}$ appeared content to let [**the dollar**]$_{\mathbf{S}}$ languish in [**a narrow range**]$_{\mathbf{X}}$ until [**tomorrow**]$_{\mathbf{X}}$, when [**the preliminary report**]$_{\mathbf{S}}$ on [**third-quarter U.S. gross national product**]$_{\mathbf{X}}$ is released.
S_3: But [**seesaw gyrations**]$_{\mathbf{S}}$ in [**the Dow Jones Industrial Average**]$_{\mathbf{X}}$ [**yesterday**]$_{\mathbf{X}}$ put [**Wall Street**]$_{\mathbf{O}}$ back in [**the spotlight**]$_{\mathbf{X}}$ and inspired [**market participants**]$_{\mathbf{O}}$ to bid [**the U.S. unit**]$_{\mathbf{S}}$ lower.

	dollar	**yesterday**	**session**	**Wall Street**	**concern**	**market**	**sessions**	**traders**	**range**	**tomorrow**	**report**	**GNP**	**gyrations**	**DJIA**	**spotlight**	**participants**
S_1	S	X	O	X	-	-	-	-	-	-	-	-	-	-	-	-
S_2	S	-	-	-	S	X	S	X	X	X	S	X	-	-	-	-
S_3	S	X	-	O	-	-	-	-	-	-	-	-	S	X	X	O

Table 1: The entity grid for the example text with three sentences and eighteen entities. Grid cells correspond to grammatical roles: subjects (S), objects (O), or neither (X).

to in d, and rows represent the sentences. Each cell corresponds to the grammatical role of an entity in the corresponding sentence: subject (S), object (O), neither (X), or nothing ($-$), and an entity is defined as a class of coreferent noun phrases. If the entity serves in multiple roles in a single sentence, then we resolve its grammatical role following the priority order: $\mathrm{S} \succ \mathrm{O} \succ \mathrm{X} \succ -$. Consider the text in our previous examples; its entity grid is shown in Table 1, and the entities are highlighted in boldface in the text above[1]. A local transition is defined as a sequence $\{\mathrm{S},\mathrm{O},\mathrm{X},-\}^n$, representing the occurrence and grammatical roles of an entity in n adjacent sentences. Such transition sequences can be extracted from the entity grid as continuous subsequences in each column. For example, the entity *dollar* in Table 1 has a bigram transition $\{\mathrm{S},\mathrm{S}\}$ from sentence 1 to 2. The entity grid is then encoded as a feature vector $\Phi(d) = (p_1(d), p_2(d), \ldots, p_m(d))$, where $p_t(d)$ is the normalized frequency of the transition t in the entity grid, and m is the number of transitions with length no more than a predefined length k. $p_t(d)$ is computed as the number of occurrences of t in the entity grid of document d, divided by the total number of transitions of the same length. Moreover, entities are differentiated by their salience — an entity is deemed to be salient if it occurs at least l times in the text, and non-salient otherwise — and transitions are computed separately for salient and non-salient entities.

3.1 Extension: Lin et al.'s Discourse Role Matrix

As mentioned previously, most extensions to B&L's entity-based local coherence model focus on enriching the feature set, including the work of Filippova and Strube (2007), Cheung and Penn (2010), Elsner and Charniak (2011), and Lin et al. (2011). To the best of our knowledge, the only exception is Feng and Hirst (2012a)'s extension from the perspective of improving the learning procedure.

Among various extensions to B&L's entity-based local coherence model, the one most related to ours is Lin et al. (2011)'s work on encoding a text as a set of entities with their associated discourse roles. Lin et al. observed that coherent texts preferentially follow certain relation patterns. However, simply using such patterns to measure the coherence of a text can result in feature sparseness. To solve this problem, they expand the relation sequence into a discourse role matrix, as shown in Table 2. Columns correspond to the entities in the text and rows represent the contiguous sentences. Each cell $\langle E_i, S_j \rangle$ corresponds to the set of discourse roles that the entity E_i serves as in sentence S_j. For example, the entity *yesterday* from S_3 takes part in Arg2 of the last relation, so the cell $\langle yesterday, S_3 \rangle$ contains the role *Contrast*.Arg2.

[1]Text elements are considered to be a single entity with multiple mentions if they refer to the same object or concept in the world, even if they have different textual realizations; e.g., *dollar* in S_1 and *U.S. unit* in S_3 refer to the same entity.

	dollar	**yesterday**	**session**	**Wall Street**	**concern**	**market**
S_1	*EntRel*.Arg1	*EntRel*.Arg1	*EntRel*.Arg1	*EntRel*.Arg1	*nil*	*nil*
S_2	*EntRel*.Arg2 *Conj*.Arg2 *Contrast*.Arg1	*nil*	*nil*	*nil*	*EntRel*.Arg2 *Conj*.Arg1 *Contrast*.Arg1	*EntRel*.Arg2 *Conj*.Arg1 *Contrast*.Arg1
S_3	*Contrast*.Arg2	*Contrast*.Arg2	*nil*	*Contrast*.Arg2	*nil*	*nil*

Table 2: A fragment of Lin et al.'s PDTB-style discourse role matrix for the example text with the first six entities across three sentences.

An entry may be empty (with a symbol *nil*, as in $\langle yesterday, S_2 \rangle$) or contain multiple discourse roles (as in $\langle dollar, S_2 \rangle$). Next, the frequencies of the discourse role transitions of lengths 2 and 3, e.g., *EntRel*.Arg1 $\rightarrow$ *Conjunction*.Arg2 and *EntRel*.Arg1 $\rightarrow$ *nil* $\rightarrow$ *Contrast*.Arg2, are calculated with respect to the matrix. For example, the frequency of *EntRel*.Arg1 $\rightarrow$ *Conjunction*.Arg2 is $1/24 = 0.042$ in Table 2.

4 Methodology

As discussed in Section 1, the main objective of our work is to study the impact of deep hierarchical discourse structures in the evaluation of text coherence. In order to conduct a direct comparison between a model with features derived from deep hierarchical discourse relations and a model with features derived from shallow discourse relations only, we adopt two separate approaches: (1) We implement a model with features derived from RST-style discourse relations, and compare it against a model with features derived from PDTB-style relations. (2) In the framework of RST-style discourse parsing, we deprive the model of any information from higher-level discourse relations and compare its performance against the model that uses the complete set of discourse relations. Moreover, as a baseline, we also re-implemented B&L's entity-based local coherence model, and we will study the effect of incorporating one of our discourse feature sets into this baseline model. Therefore, we have four ways to encode discourse relation features, namely, entity-based, PDTB-style, full RST-style, and shallow RST-style.

4.1 Entity-based Feature Encoding

In entity-based feature encoding, our goal is to formulate a text into an entity grid, such as the one shown in Table 1, from which we extract entity-based local transitions. In our re-implementation of B&L, we use the same parameter settings as B&L's original model, i.e., the optimal transition length $k = 3$ and the salience threshold $l = 2$. However, when extracting entities in each sentence, e.g., *dollar*, *yesterday*, etc., we do not perform coreference resolution; rather, for better coverage, we follow the suggestion of Elsner and Charniak (2011) and extract all nouns (including non-head nouns) as entities. We use the Stanford dependency parser (de Marneffe et al., 2006) to extract nouns and their grammatical roles. This strategy of entity extraction also applies to the other three feature encoding methods to be described below.

4.2 PDTB-style Feature Encoding

To encode PDTB-style discourse relations into the model, we parse the texts using an end-to-end PDTB-style discourse parser[2] developed by Lin et al. (2014). The F_1 score of this parser is around 85% for recognizing explicit relations and around 40% for recognizing implicit relations. A text is thus represented by a discourse role matrix in the same way as shown in Table 2. Most parameters in our PDTB-style feature encoding follow those of Lin et al. (2011): each entity is associated with the fully-fledged discourse roles, i.e., with type and argument information included; the maximum length of discourse role transitions is 3; and transitions are generated separately for salient and non-salient entities with a threshold set at 2. However, compared to Lin et al.'s model, there are two differences in our re-implementation, and evaluated on a held-out development set, these modifications are shown to be effective in improving the performance.

[2] `http://wing.comp.nus.edu.sg/~linzihen/parser/`

	dollar	yesterday	session	Wall Street	concern	market
S_1	*Background*.N *Temporal*.N	*Background*.N *Temporal*.N	*Temporal*.S	*Temporal*.S	*nil*	*nil*
S_2	*List*.N *Condition*.N *Contrast*.S	*nil*	*nil*	*nil*	*List*.N *Contrast*.S	*List*.N *Contrast*.S
S_3	*Contrast*.N *Background*.N *Cause*.N	*Cause*.S	*nil*	*Cause*.S	*nil*	*nil*

Table 3: A fragment of the full RST-style discourse role matrix for the example text with the first six entities across three sentences.

First, we differentiate between intra- and multi-sentential discourse relations, which is motivated by a finding in the field of RST-style discourse parsing — distributions of various discourse relation types are quite distinct between intra-sentential and multi-sentential instances (Feng and Hirst, 2012b; Joty et al., 2012) — and we assume that a similar phenomenon exists for PDTB-style discourse relations. Therefore, we assign two sets of discourse roles to each entity: intra-sentential and multi-sentential roles, which are the roles that the entity plays in the corresponding intra- and multi-sentential relations.

Second, instead of Level-1 PDTB discourse relations (6 in total), we use Level-2 relations (18 in total) in feature encoding, so that richer information can be captured in the model, resulting in $18 \times 2 = 36$ different discourse roles with argument attached. We then generate four separate set of features for the combination of intra-/multi-sentential discourse relation roles, and salient/non-salient entities, among which transitions consisting of only *nil* symbols are excluded. Therefore, the total number of features in PDTB-style encoding is $4 \times (36^2 + 36^3 - 2) \approx 192$K.

4.3 Full RST-style Feature Encoding

For RST-style feature encoding, we parse the texts using an end-to-end RST-style discourse parser developed by Feng and Hirst (2014), which produces a discourse tree representation for each text, such as the one shown in Figure 2. For relation labeling, the overall accuracy of this discourse parser is 58%, evaluated on the RST-DT.

We encode the RST-style discourse relations in a similar fashion to PDTB-style encoding. However, since the definition of discourse roles depends on the particular discourse framework, here, we adapt Lin et al.'s PDTB-style encoding by replacing the PDTB-style discourse relations with RST-style discourse relations, and the argument information (Arg1 or Arg2) by the nuclearity information (nucleus or the satellite) in an RST-style discourse relation. More importantly, in order to reflect the hierarchical structure in an RST-style discourse parse tree, when extracting the set of discourse relations that an entity participates in, we find all those discourse relations that the entity appears in the main EDUs of each relation[3] and represent the role of the entity in each of these discourse relations. In this way, we can encode long-distance discourse relations for the most relevant entities. For example, considering the RST-style discourse tree representation in Figure 2, we encode the *Background* relation for the entities *dollar* and *yesterday* in S_1, as well as the entity *dollar* in S_3, but not for the remaining entities in the text, even though the *Background* relation covers the whole text. The corresponding full RST-style discourse role matrix for the example text is shown in Table 3.

As in PDTB-style feature encoding, we differentiate between intra- and multi-sentential discourse relations; we use 17 coarse-grained classes of RST-style relations in feature encoding; the optimal transi-

[3]The main EDUs of a discourse relation are the EDUs obtained by traversing the discourse subtree in which the relation of interest constitutes the root node, following the nucleus branches down to the leaves. For instance, for the RST discourse tree in Figure 2, the main EDUs of the *Background* relation on the top level are $\{e_1, e_7\}$, and the main EDUs of the *List* relation among (e_3-e_5) are $\{e_3, e_4\}$.

tion length k is 3; and the salience threshold l is 2. The total number of features in RST-style encoding is therefore $4 \times (34^2 + 34^3 - 2) \approx 162$K, which is roughly the same as that in PDTB-style feature encoding.

4.4 Shallow RST-style Feature Encoding

Shallow RST-style encoding is almost identical to full RST-style encoding, as introduced in Section 4.3, except that, when we derive discourse roles, we consider shallow discourse relations only. To be consistent with the majority of PDTB-style discourse relations, we define shallow discourse relations as those relations which hold between text spans of the same sentence, or between two adjacent sentences. For example, in Figure 2, the *Background* relation between (e_1-e_2) and (e_3-e_7) is not a shallow discourse relation (it holds between a single sentence and the concatenation of two sentences), and thus will be excluded from shallow RST-style feature encoding.

5 Experiments

To evaluate our proposed model with deep discourse structures encoded, we conduct two series of experiments on two different datasets, each of which simulates a sub-task in the evaluation of text coherence, i.e., **sentence ordering** and **essay scoring**. Since text coherence is a matter of degree rather than a binary classification, in both evaluation tasks we formulate the problem as a pairwise preference ranking problem. Specifically, given a set of texts with different degrees of coherence, we train a ranker which learns to prefer a more coherent text over a less coherent counterpart. Accuracy is therefore measured as the fraction of correct pairwise rankings as recognized by the ranker. In our experiments, we use the SVM^{light} package[4] (Joachims, 1999) with the ranking configuration, and all parameters are set to their default values.

5.1 Sentence Ordering

The task of sentence ordering, which has been extensively studied in previous work, attempts to simulate the situation where, given a predefined set of information-bearing items, we need to determine the best order in which the items should be presented. As argued by Barzilay and Lapata (2005), sentence ordering is an essential step in many content-generation components, such as multi-document summarization.

In this task, we use a dataset consisting of a subset of the Wall Street Journal (WSJ) corpus, in which the minimum length of a text is 20 sentences, and the average length is 41 sentences. For each text, we create 20 random permutations by shuffling the original order of the sentences. In total, we have 735 source documents and $735 \times 20 = 14,700$ permutations. Because the RST-style discourse parser we use is trained on a fraction of the WSJ corpus, we remove the training texts from our dataset, to guarantee that the discourse parser will not perform exceptionally well on some particular texts. However, since the PDTB-style discourse parser we use is trained on almost the entire WSJ corpus, we cannot do the same for the PDTB-style parser.

In this experiment, our learning instances are pairwise ranking preferences between a source text and one of its permutations, where the source text is always considered more coherent than its permutations. Therefore, we have $735 \times 20 = 14,700$ total pairwise rankings, and we conduct 5-fold cross-validation on five disjoint subsets. In each fold, one-fifth of the rankings are used for testing, and the rest for training.

5.2 Essay Scoring

The second task is essay scoring, and we use a subset of International Corpus of Learner English (ICLE) (Granger et al., 2009). The dataset consists of 1,003 essays about 34 distinct topics, written by university undergraduates speaking 14 native languages who are learners of English as a Foreign Language. Each essay has been annotated with an organization score from 1 to 4 at half-point increments by Persing et al. (2010). We use these organization scores to *approximate* the degrees of coherence in the essays. The average length of the essays is 32 sentences, and the average organization score is 3.05, with a standard deviation of 0.59.

[4] `http://svmlight.joachis.org/`

Model		sentence ordering	essay scoring
No discourse structure	Entity	95.1	66.4
Shallow discourse structures	PDTB	97.2	82.2
	PDTB&Entity	97.3	83.3
	Shallow RST	98.5	87.2
	Shallow RST&Entity	98.8	87.2
Deep discourse structures	Full RST	99.1	**88.3**
	Full RST&Entity	**99.3**	87.7

Table 4: Accuracy (%) of various models on the two evaluation tasks: sentence ordering and essay scoring. For sentence ordering, accuracy difference is significant with $p < .01$ for all pairs of models except between PDTB and PDTB&Entity. For essay scoring, accuracy difference is significant with $p < .01$ for all pairs of models except between shallow RST and shallow RST&Entity. Significance is determined with the Wilcoxon signed-rank test.

In this experiment, our learning instances are pairwise ranking preferences between a pair of essays on the same topic written by students speaking the same native language, excluding pairs with the same organization score. In total, we have 22,362 pairwise rankings. Similarly, we conduct 5-fold cross-validations on these rankings.

In fact, the two datasets used in the two evaluation tasks reflect different characteristics by themselves. The WSJ dataset, although somewhat artificial due to the permuting procedure, is representative of texts with well-formed syntax. By contrast, the ICLE dataset, although not artificial, contains occasional syntactic errors, because the texts are written by non-native English speakers. Therefore, using these two distinct datasets allows us to evaluate our models in tasks where different challenges may be expected.

6 Results

In this section, we demonstrate the performance of our models with discourse roles encoded in one of the three ways: PDTB-style, full RST-style or shallow RST-style, and compare against their combination with our re-implemented B&L's entity-based local transition features. The evaluation is conducted on the two tasks, sentence ordering and essay scoring, and the accuracy is reported as the fraction of correct pairwise rankings averaged over 5-fold cross-validation.

The performance of various models is shown in Table 4. The first section of the table shows the results of our re-implementation of B&L's entity-based local coherence model, representing the effect with **no** discourse structure encoded. The second section shows the results of four models with **shallow** discourse structures encoded, including the two basic models, PDTB-style and shallow RST-style feature encoding, and their combination with the entity-based feature encoding. The last section shows the results of our models with **deep** discourse structures encoded, including the RST-style feature encoding and its combination with the entity-base feature encoding. With respect to the performance, we observe a number of consistent patterns across both evaluation tasks.

First, with **no** discourse structure encoded, the entity-based model (the first row) performs the worst among all models, suggesting that discourse structures are truly important and can capture coherence in a more sophisticated way than pure grammatical roles. Moreover, the performance gap is particularly large for essay scoring, which is probably due to the fact that, as argued by Persing et al. (2010), the organization score, which we use to approximate the degrees of coherence, is not equivalent to text coherence. Organization relates more to the logical development in the texts, while coherence is about lexical and semantic continuity; but discourse relations can capture the logical relations at least to some extent.

Secondly, with **deep** discourse structures encoded, the RST-style model in the third section significantly outperforms ($p < .01$) the models with shallow discourse structures, i.e., the PDTB-style and

shallow RST-style models in the middle section, confirming our intuition that deep discourse structures are more powerful than shallow structures. This is also the case when entity-based features are included.

Finally, considering the models in the middle section of the table, we can gain more insight into the difference between PDTB-style and RST-style encoding. As can be seen, even without information from the more powerful deep hierarchical discourse structures, shallow RST-style encoding still significantly outperforms PDTB-style encoding on both tasks ($p < .01$). This is primarily due to the fact that the discourse relations discovered by RST-style parsing have wider coverage of the text[5], and thus induce richer information about the text. Therefore, because of its ability to annotate deep discourse structures and its better coverage of discourse relations, RST-style discourse parsing is generally more powerful than PDTB-style parsing, as far as coherence evaluation is concerned.

However, with respect to combining full RST-style features with entity features, we have contradictory results on the two tasks: for sentence ordering, the combination is significantly better than each single model, while for essay scoring, the combination is worse than using RST-style features alone. This is probably related to the previously discussed issue of using entity-based features for essay scoring, due to the subtle difference between coherence and organization.

7 Conclusion and Future Work

In this paper, we have studied the impact of deep discourse structures in the evaluation of text coherence by two approaches. In the first approach, we implemented a model with discourse role features derived from RST-style discourse parsing, which represents deep discourse structures, and compared it against our re-implemented Lin et al. (2011)'s model derived from PDTB-style parsing, with no deep discourse structures annotated. In the second approach, we compared our complete RST-style model against a model with shallow RST-style encoding. Evaluated on the two tasks, sentence ordering and essay scoring, deep discourse structures are shown to be effective for better differentiation of text coherence. Moreover, we showed that, even without deep discourse structures, shallow RST-style encoding is more powerful than PDTB-style encoding, because it has better coverage of discourse relations in texts. Finally, combining discourse relations with entity-based features is shown to have an inconsistent effect on the two evaluation tasks, which is probably due to the different nature of the two tasks.

In our future work, we wish to explore the effect of automatic discourse parsers in our methodology. As discussed previously, the PDTB- and RST-style discourse parsers used in our experiments are far from perfect. Therefore, it is possible that using automatically extracted discourse relations creates some bias to the training procedure; it is also possible that what our model actually learns is the distribution over those discourse relations which automatic discourse parsers are mostly confident with, and thus errors (if any) made on other relations do not matter. One potential way to verify these two possibilities is to study the effect of each particular type of discourse relation to the resulting model, and we leave it for future exploration.

Acknowledgements

We thank the reviewers for their valuable advice and comments. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada and by the University of Toronto.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 42rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 141–148.

Jackie Chi Kit Cheung and Gerald Penn. 2010. Entity-based local coherence modelling using topological fields. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 186–195.

[5]The entire text is covered by the annotation produced by RST-style discourse parsing, while this is generally not true for PDTB-style discourse parsing.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 125–129.

Vanessa Wei Feng and Graeme Hirst. 2012a. Extending the entity-based coherence model with multiple ranks. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 315–324, Avignon, France.

Vanessa Wei Feng and Graeme Hirst. 2012b. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, pages 60–68, Jeju, Korea.

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.

Katja Filippova and Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 2007)*, pages 139–142.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English (Version 2)*. Presses universitaires de Louvain.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL 2012, pages 904–915.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011)*, Portland, Oregon, USA, June.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 2:151–184.

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

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA, October. Association for Computational Linguistics.

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 6th International Conference on Language Resources and Evaluation (LREC 2008)*.