Automatically Evaluating Text Coherence Using Discourse Relations

Ziheng Lin, Hwee Tou Ng and **Min-Yen Kan**
Department of Computer Science
National University of Singapore
13 Computing Drive
Singapore 117417
{linzihen,nght,kanmy}@comp.nus.edu.sg

Abstract

We present a novel model to represent and assess the discourse coherence of text. Our model assumes that coherent text implicitly favors certain types of discourse relation transitions. We implement this model and apply it towards the text ordering ranking task, which aims to discern an original text from a permuted ordering of its sentences. The experimental results demonstrate that our model is able to significantly outperform the state-of-the-art coherence model by Barzilay and Lapata (2005), reducing the error rate of the previous approach by an average of 29% over three data sets against human upper bounds. We further show that our model is synergistic with the previous approach, demonstrating an error reduction of 73% when the features from both models are combined for the task.

1 Introduction

The coherence of a text is usually reflected by its discourse structure and relations. In Rhetorical Structure Theory (RST), Mann and Thompson (1988) observed that certain RST relations tend to favor one of two possible canonical orderings. Some relations (*e.g.*, Concessive and Conditional) favor arranging their satellite span before the nucleus span. In contrast, other relations (*e.g.*, Elaboration and Evidence) usually order their nucleus before the satellite. If a text that uses non-canonical relation orderings is rewritten to use canonical orderings, it often improves text quality and coherence.

This notion of preferential ordering of discourse relations is observed in natural language in general, and generalizes to other discourse frameworks aside from RST. The following example shows a Contrast relation between the two sentences.

(1) [Everyone agrees that most of the nation's old bridges need to be repaired or replaced. $]_{S_1}$ [But there's disagreement over how to do it. $]_{S_2}$

Here the second sentence provides contrasting information to the first. If this order is violated without rewording (*i.e.*, if the two sentences are swapped), it produces an incoherent text (Marcu, 1996).

In addition to the intra-relation ordering, such preferences also extend to inter-relation ordering:

(2) [The Constitution does not expressly give the president such power. $]_{S_1}$ [However, the president does have a duty not to violate the Constitution. $]_{S_2}$ [The question is whether his only means of defense is the veto. $]_{S_3}$

The second sentence above provides a contrast to the previous sentence and an explanation for the next one. This pattern of Contrast-followed-by-Cause is rather common in text (Pitler et al., 2008). Ordering the three sentences differently results in incoherent, cryptic text.

Thus coherent text exhibits measurable preferences for specific intra- and inter-discourse relation ordering. Our key idea is to use the converse of this phenomenon to assess the coherence of a text. In this paper, we detail our model to capture the coherence of a text based on the statistical distribution of the discourse structure and relations. Our method specifically focuses on the discourse relation transitions between adjacent sentences, modeling them in a discourse role matrix.

Our study makes additional contributions. We implement and validate our model on three data sets, which show robust improvements over the current state-of-the-art for coherence assessment. We also provide the first assessment of the upper-bound of human performance on the standard task of distinguishing coherent from incoherent orderings. To the best our knowledge, this is also the first study in which we show output from an automatic discourse parser helps in coherence modeling.

2 Related Work

The study of coherence in discourse has led to many linguistic theories, of which we only discuss algorithms that have been reduced to practice.

Barzilay and Lapata (2005; 2008) proposed an entity-based model to represent and assess *local* textual coherence. The model is motivated by Centering Theory (Grosz et al., 1995), which states that subsequent sentences in a locally coherent text are likely to continue to focus on the same entities as in previous sentences. Barzilay and Lapata operationalized Centering Theory by creating an entity grid model to capture discourse entity transitions at the sentence-to-sentence level, and demonstrated their model's ability to discern coherent texts from incoherent ones. Barzilay and Lee (2004) proposed a domain-dependent HMM model to capture topic shift in a text, where topics are represented by hidden states and sentences are observations. The *global* coherence of a text can then be summarized by the overall probability of topic shift from the first sentence to the last. Following these two directions, Soricut and Marcu (2006) and Elsner et al. (2007) combined the entity-based and HMM-based models and demonstrated that these two models are complementary to each other in coherence assessment.

Our approach differs from these models in that it introduces and operationalizes another indicator of discourse coherence, by modeling a text's discourse relation transitions. Karamanis (2007) has tried to integrate local discourse relations into the Centering-based coherence metrics for the task of information ordering, but was not able to obtain improvement over the baseline method, which is partly due to the much smaller data set and the way the discourse relation information is utilized in heuristic constraints and rules.

To implement our proposal, we need to identify the text's discourse relations. This task, *discourse parsing*, has been a recent focus of study in the natural language processing (NLP) community, largely enabled by the availability of large-scale discourse annotated corpora (Wellner and Pustejovsky, 2007; Elwell and Baldridge, 2008; Lin et al., 2009; Pitler et al., 2009; Pitler and Nenkova, 2009; Lin et al., 2010; Wang et al., 2010). The Penn Discourse Treebank (PDTB) (Prasad et al., 2008) is such a corpus which provides a discourse-level annotation on top of the Penn Treebank, following a predicate-argument approach (Webber, 2004). Crucially, the PDTB provides annotations not only on explicit (*i.e.*, signaled by discourse connectives such as *because*) discourse relations, but also implicit (*i.e.*, inferred by readers) ones.

3 Using Discourse Relations

To utilize discourse relations of a text, we first apply automatic discourse parsing on the input text. While any discourse framework, such as the Rhetorical Structure Theory (RST), could be applied in our work to encode discourse information, we have chosen to work with the Discourse Lexicalized Tree Adjoining Grammar (D-LTAG) by Webber (2004) as embodied in the PDTB, as a PDTB-styled discourse parser[1] developed by Lin et al. (2010) has recently become freely available.

This parser tags each explicit/implicit relation with two levels of relation types. In this work, we utilize the four PDTB Level-1 types: Temporal (Temp), Contingency (Cont), Comparison (Comp), and Expansion (Exp). This parser automatically identifies the discourse relations, labels the argument spans, and classifies the relation types, including identifying common entity and no relation (EntRel and NoRel) as types.

A simple approach to directly model the connections among discourse relations is to use the sequence of discourse relation transitions. Text (2) in Section 1 can be represented by $S_1 \xrightarrow{Comp} S_2 \xrightarrow{Cont} S_3$, for instance, when we use Level-1 types. In such a basic approach, we can compile a distribu-

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

tion of the n-gram discourse relation transition sequences in gold standard coherent text, and a similar one for incoherent text. For example, the above text would generate the transition bigram Comp→Cont. We can build a classifier to distinguish one from the other through learned examples or using a suitable distribution distance measure (*e.g.*, KL Divergence).

In our pilot work where we implemented such a basic model with n-gram features for relation transitions, the performance was very poor. Our analysis revealed a serious shortcoming: as the discourse relation transitions in short texts are few in number, we have very little data to base the coherence judgment on. However, when faced with even short text excerpts, humans can distinguish coherent texts from incoherent ones, as exemplified in our example texts. The basic approach also does not model the intra-relation preference. In Text (1), a Comparison (Comp) relation would be recorded between the two sentences, irregardless of whether S_1 or S_2 comes first. However, it is clear that the ordering of ($S_1 \prec S_2$) is more coherent.

4 A Refined Approach

The central problem with the basic approach is in its sparse modeling of discourse relations. In developing an improved model, we need to better exploit the discourse parser's output to provide more circumstantial evidence to support the system's coherence decision.

In this section, we introduce the concept of a discourse role matrix which aims to capture an expanded set of discourse relation transition patterns. We describe how to represent the coherence of a text with its discourse relations and how to transform such information into a matrix representation. We then illustrate how we use the matrix to formulate a preference ranking problem.

4.1 Discourse Role Matrix

Figure 1 shows a text and its gold standard PDTB discourse relations. When a term appears in a discourse relation, the discourse role of this term is defined as the discourse relation type plus the argument span in which the term is located (*i.e.*, the argument tag). For instance, consider the term "cananea" in the first relation. Since the relation type is a Comparison and "cananea" is found in the Arg1 span, the discourse role of "cananea" is defined as Comp.Arg1. When terms appear in different relations and/or argument spans, they obtain different discourse roles in the text. For instance, "cananea" plays a different discourse role of Temp.Arg1 in the third relation in Figure 1. In the fourth relation, since "cananea" appears in both argument spans, it has two additional discourse roles, Exp.Arg1 and

[Japan normally depends heavily on the Highland Valley and **Cananea** mines as well as the Bougainville mine in Papua New Guinea. $]_{S_1}$ [Recently, Japan has been buying copper elsewhere. $]_{S_2}$ [[But as Highland Valley and **Cananea** begin operating, $]_{C_{3.1}}$ [they are expected to resume their roles as Japan's suppliers. $]_{C_{3.2}}$ $]_{S_3}$ [[According to Fred Demler, metals economist for Drexel Burnham Lambert, New York, $]_{C_{4.1}}$ ["Highland Valley has already started operating $]_{C_{4.2}}$ [and **Cananea** is expected to do so soon." $]_{C_{4.3}}$ $]_{S_4}$

5 discourse relations are present in the above text:

1. Implicit Comparison between S_1 as Arg1, and S_2 as Arg2
2. Explicit Comparison using "but" between S_2 as Arg1, and S_3 as Arg2
3. Explicit Temporal using "as" within S_3 (Clause $C_{3.1}$ as Arg1, and $C_{3.2}$ as Arg2)
4. Implicit Expansion between S_3 as Arg1, and S_4 as Arg2
5. Explicit Expansion using "and" within S_4 (Clause $C_{4.2}$ as Arg1, and $C_{4.3}$ as Arg2)

Figure 1: An excerpt with four contiguous sentences from wsj_0437, showing five gold standard discourse relations. "Cananea" is highlighted for illustration.

S#	Terms				
	copper	cananea	operat	depend	...
S_1	nil	Comp.Arg1	nil	Comp.Arg1	
S_2	Comp.Arg2 Comp.Arg1	nil	nil	nil	
S_3	nil	Comp.Arg2 Temp.Arg1 Exp.Arg1	Comp.Arg2 Temp.Arg1 Exp.Arg1	nil	
S_4	nil	Exp.Arg2	Exp.Arg1 Exp.Arg2	nil	

Table 1: Discourse role matrix fragment for Figure 1. Rows correspond to sentences, columns to stemmed terms, and cells contain extracted discourse roles.

Exp.Arg2. The discourse role matrix thus represents the different discourse roles of the terms across the continuous text units. We use sentences as the text units, and define terms to be the stemmed forms of the open class words: nouns, verbs, adjectives, and adverbs. We formulate the discourse role matrix such that it encodes the discourse roles of the terms across adjacent sentences.

Table 1 shows a fragment of the matrix representation of the text in Figure 1. Columns correspond to the extracted terms; rows, the contiguous sentences. A cell C_{T_i,S_j} then contains the set of the discourse roles of the term T_i that appears in sentence S_j. For example, the term "cananea" from S_1 takes part in the first relation, so the cell $C_{cananea,S_1}$ contains the role Comp.Arg1. A cell may be empty (*nil*, as in $C_{cananea,S_2}$) or contain multiple discourse roles (as in $C_{cananea,S_3}$, as "cananea" in S_3 participates in the second, third, and fourth relations). Given these discourse relations, building the matrix is straightforward: we note down the relations that a term T_i from a sentence S_j participates in, and record its discourse roles in the respective cell.

We hypothesize that the sequence of discourse role transitions in a coherent text provides clues that distinguish it from an incoherent text. The discourse role matrix thus provides the foundation for computing such role transitions, on a per term basis. In fact, each column of the matrix corresponds to a lexical chain (Morris and Hirst, 1991) for a particular term across the whole text. The key differences from the traditional lexical chains are that our chain nodes' entities are simplified (they share the same stemmed form, instead being connected by WordNet relations), but are further enriched by being typed with discourse relations.

We compile the set of sub-sequences of discourse role transitions for every term in the matrix. These transitions tell us how the discourse role of a term varies through the progression of the text. For instance, "cananea" functions as Comp.Arg1 in S_1 and Comp.Arg2 in S_3, and plays the role of Exp.Arg1 and Exp.Arg2 in S_3 and S_4, respectively. As we have six relation types (Temp(oral), Cont(ingency), Comp(arison), Exp(ansion), EntRel and NoRel) and two argument tags (Arg1 and Arg2) for each type, we have a total of $6 \times 2 = 12$ possible discourse roles, plus a *nil* value. We define a *discourse role transition* as the sub-sequence of discourse roles for a term in multiple consecutive sentences. For example, the discourse role transition of "cananea" from S_1 to S_2 is Comp.Arg1→nil. As a cell may contain multiple discourse roles, a transition may produce multiple sub-sequences. For example, the length 2 sub-sequences for "cananea" from S_3 to S_4, are Comp.Arg2→Exp.Arg2, Temp.Arg1→Exp.Arg2, and Exp.Arg1→Exp.Arg2.

Each sub-sequence has a probability that can be computed from the matrix. To illustrate the calculation, suppose the matrix fragment in Table 1 is the entire discourse role matrix. Then since there are in total 25 length 2 sub-sequences and the sub-sequence Comp.Arg2→Exp.Arg2 has a count of two, its probability is $2/25 = 0.08$. A key property of our approach is that, while discourse transitions are captured locally on a per-term basis, the probabilities of the discourse transitions are aggregated globally, across all terms. We believe that the overall distribution of discourse role transitions for a coherent text is distinguishable from that for an incoherent text. Our model captures the distributional differences of such sub-sequences in coherent and incoherent text in training to determine an unseen text's coherence. To evaluate the coherence of a text, we extract sub-sequences with various lengths from the discourse role matrix as features[2] and compute the sub-sequence probabilities as the feature values.

To further refine the computation of the sub-sequence distribution, we follow (Barzilay and Lapata, 2005) and divide the matrix into a salient matrix and a non-salient matrix. Terms (columns) with a frequency greater than a threshold form the salient matrix, while the rest form the non-salient matrix. The sub-sequence distributions are then calculated separately for these two matrices.

4.2 Preference Ranking

While some texts can be said to be simply coherent or incoherent, often it is a matter of degree. A text can be less coherent when compared to one text, but more coherent when compared to another. As such, since the notion of coherence is relative, we feel that coherence assessment is better represented as

[2] Sub-sequences consisting of only *nil* values are not used as features.

a ranking problem rather than a classification problem. Given a pair of texts, the system ranks them based on how coherent they are. Applications of such a system include differentiating a text from its permutation (*i.e.*, the sentence ordering of the text is shuffled) and identifying a more well-written essay from a pair. Such a system can easily generalize from pairwise ranking into listwise, suitable for the ordinal ranking of a set of texts. Coherence scoring equations can also be deduced (Lapata and Barzilay, 2005) from such a model, yielding coherence scores.

To induce a model for preference ranking, we use the SVM^{light} package[3] by (Joachims, 1999) with the preference ranking configuration for training and testing. All parameters are set to their default values.

5 Experiments

We evaluate our coherence model on the task of *text ordering ranking*, a standard coherence evaluation task used in both (Barzilay and Lapata, 2005) and (Elsner et al., 2007). In this task, the system is asked to decide which of two texts is more coherent. The pair of texts consists of a source text and one of its permutations (*i.e.*, the text's sentence order is randomized). Assuming that the original text is always more discourse-coherent than its permutation, an ideal system will prefer the original to the permuted text. A system's accuracy is thus the number of times the system correctly chooses the original divided by the total number of test pairs.

In order to acquire a large data set for training and testing, we follow the approach in (Barzilay and Lapata, 2005) to create a collection of synthetic data from *Wall Street Journal* (WSJ) articles in the Penn Treebank. All of the WSJ articles are randomly split into a training and a testing set; 40 articles are held out from the training set for development. For each article, its sentences are permuted up to 20 times to create a set of permutations[4]. Each permutation is paired with its source text to form a pair.

We also evaluate on two other data collections (cf. Table 2), provided by (Barzilay and Lapata, 2005), for a direct comparison with their entity-based model. These two data sets consist of Associated Press articles about earthquakes from the North American News Corpus, and narratives from the National Transportation Safety Board. These collections are much smaller than the WSJ data, as each training/testing set contains only up to 100 source articles. Similar to the WSJ data, we construct pairs by permuting each source article up to 20 times.

		WSJ	Earthquakes	Accidents
Train	# Articles	1040	97	100
	# Pairs	19120	1862	1996
	Avg. # Sents	22.0	10.4	11.5
Test	# Articles	1079	99	100
	# Pairs	19896	1956	1986

Table 2: Details of the WSJ, Earthquakes, and Accidents data sets, showing the number of training/testing articles, number of pairs of articles, and average length of an article (in sentences).

Our model has two parameters: (1) the term frequency (TF) that is used as a threshold to identify salient terms, and (2) the lengths of the sub-sequences that are extracted as features. These parameters are tuned on the development set, and the best ones that produce the optimal accuracy are TF $>=$ 2 and lengths of the sub-sequences $<=$ 3.

We must also be careful in using the automatic discourse parser. We note that the discourse parser of Lin et al. (2010) comes trained on the PDTB, which provides annotations on top of the whole WSJ data. As we also use the WSJ data for evaluation, we must avoid parsing an article that has already been used in training the parser to prevent training on the test data. We re-train the parser with 24 WSJ sections and use the trained parser to parse the sentences in our WSJ collection from the remaining section. We repeat this re-training/parsing process for all 25 sections. Because the Earthquakes and Accidents data do not overlap with the WSJ training data, we use the parser as distributed to parse these two data sets. Since the discourse parser utilizes paragraph boundaries but a permuted text does not have such boundaries, we ignore paragraph boundaries and treat the source text as if it has only one paragraph. This is to make sure that we do not give the system extra information because of this difference between the source and permuted text.

[3] `http://svmlight.joachims.org/`

[4] Short articles may produce less than 20 permutations.

5.1 Human Evaluation

While the text ordering ranking task has been used in previous studies, two key questions about this task have remained unaddressed in the previous work: (1) to what extent is the assumption that the source text is more coherent than its permutation correct? and (2) how well do humans perform on this task? The answer to the first is needed to validate the correctness of this synthetic task, while the second aims to obtain the upper bound for evaluation. We conduct a human evaluation to answer these questions.

We randomly select 50 source text/permutation pairs from each of the WSJ, Earthquakes, and Accidents training sets. We observe that some of the source texts have formulaic structures in their initial sentences that give away the correct ordering. Sources from the Earthquakes data always begin with a headline sentence and a location-newswire sentence, and many sources from the Accidents data start with two sentences of "This is preliminary . . . errors. Any errors . . . completed." We remove these sentences from the source and permuted texts, to avoid the subjects judging based on these clues instead of textual coherence. For each set of 50 pairs, we assigned two human subjects (who are not authors of this paper) to perform the ranking. The subjects are told to identify the source text from the pair. When both subjects rank a source text higher than its permutation, we interpret it as the subjects agreeing that the source text is more coherent than the permutation. Table 3 shows the inter-subject agreements.

WSJ	Earthquakes	Accidents	Overall
90.0	90.0	94.0	91.3

Table 3: Inter-subject agreements on the three data sets.

While our study is limited and only indicative, we conclude from these results that the task is tractable. Also, since our subjects' judgments correlate highly with the gold standard, the assumption that the original text is always more coherent than the permuted text is supported. Importantly though, human performance is not perfect, suggesting fair upper bound limits on system performance. We note that the Accidents data set is relatively easier to rank, as it has a higher upper bound than the other two.

5.2 Baseline

Barzilay and Lapata (2005) showed that their entity-based model is able to distinguish a source text from its permutation accurately. Thus, it can serve as a good comparison point for our discourse relation-based model. We compare against their Syntax+Salience setting. Since they did not automatically determine the coreferential information of a permuted text but obtained that from its corresponding source text, we do not perform automatic coreference resolution in our reimplementation of their system. For fair comparison, we follow their experiment settings as closely as possible. We re-use their Earthquakes and Accidents dataset as is, using their exact permutations and pre-processing. For the WSJ data, we need to perform our own pre-processing, thus we employed the Stanford parser[5] to perform sentence segmentation and constituent parsing, followed by entity extraction.

5.3 Results

We perform a series of experiments to answer the following four questions:

1. Does our model outperform the baseline?
2. How do the different features derived from using relation types, argument tags, and salience information affect performance?
3. Can the combination of the baseline and our model outperform the single models?
4. How does system performance of these models compare with human performance on the task?

Baseline results are shown in the first row of Table 4. The results on the Earthquakes and Accidents data are quite similar to those published in (Barzilay and Lapata, 2005) (they reported 83.4% on Earthquakes and 89.7% on Accidents), validating the correctness of our reimplementation of their method.

Row 2 in Table 4 shows the overall performance of the proposed refined model, answering Question 1. The model setting of Type+Arg+Sal means that the model makes use of the discourse roles consisting of 1) relation types and 2) argument tags (*e.g.*,

[5] `http://nlp.stanford.edu/software/lex-parser.shtml`

	WSJ	Earthquakes	Accidents
Baseline	85.71	83.59	89.93
Type+Arg+Sal	**88.06****	**86.50****	**89.38**
Arg+Sal	88.28**	85.89*	87.06
Type+Sal	87.06**	82.98	86.05
Type+Arg	85.98	82.67	87.87
Baseline & Type+Arg+Sal	89.25**	89.72**	91.64**

Table 4: Test set ranking accuracy. The first row shows the baseline performance, the next four show our model with different settings, and the last row is a combined model. Double (**) and single (*) asterisks indicate that the respective model significantly outperforms the baseline at $p < 0.01$ and $p < 0.05$, respectively. We follow Barzilay and Lapata (2008) and use the Fisher Sign test.

the discourse role Comp.Arg2 consists of the type Comp(arison) and the tag Arg2), and 3) two distinct feature sets from salient and non-salient terms. Comparing these accuracies to the baseline, our model significantly outperforms the baseline with $p < 0.01$ in the WSJ and Earthquakes data sets with accuracy increments of 2.35% and 2.91%, respectively. In Accidents, our model's performance is slightly lower than the baseline, but the difference is not statistically significant.

To answer Question 2, we perform feature ablation testing. We eliminate each of the information sources from the full model. In **Row 3**, we first delete relation types from the discourse roles, which causes discourse roles to only contain the argument tags. A discourse role such as Comp.Arg2 becomes Arg2 after deleting the relation type. Comparing Row 3 to Row 2, we see performance reductions on the Earthquakes and Accidents data after eliminating type information. **Row 4** measures the effect of omitting argument tags (Type+Sal). In this setting, the discourse role Comp.Arg2 reduces to Comp. We see a large reduction in performance across all three data sets. This model is also most similar to the basic naïve model in Section 3. These results suggest that the argument tag information plays an important role in our discourse role transition model. **Row 5** omits the salience information (Type+Arg), which also markedly reduces performance. This result supports the use of salience, in line with the conclusion drawn in (Barzilay and Lapata, 2005).

To answer Question 3, we train and test a combined model using features from both the baseline and our model (shown as **Row 6** in Table 4). The entity-based model of Barzilay and Lapata (2005) connects the local entity transition with textual coherence, while our model looks at the patterns of discourse relation transitions. As these two models focus on different aspects of coherence, we expect that they are complementary to each other. The combined model in all three data sets gives the highest performance in comparison to all single models, and it significantly outperforms the baseline model with $p < 0.01$. This confirms that the combined model is linguistically richer than the single models as it integrates different information together, and the entity-based model and our model are synergistic.

To answer Question 4, when compared to the human upper bound (Table 3), the performance gaps for the baseline model are relatively large, while those for our full model are more acceptable in the WSJ and Earthquakes data. For the combined model, the error rates are significantly reduced in all three data sets. The average error rate reductions against 100% are 9.57% for the full model and 26.37% for the combined model. If we compute the average error rate reductions against the human upper bounds (rather than an oracular 100%), the average error rate reduction for the full model is 29% and that for the combined model is 73%. While these are only indicative results, they do highlight the significant gains that our model is making towards reaching human performance levels.

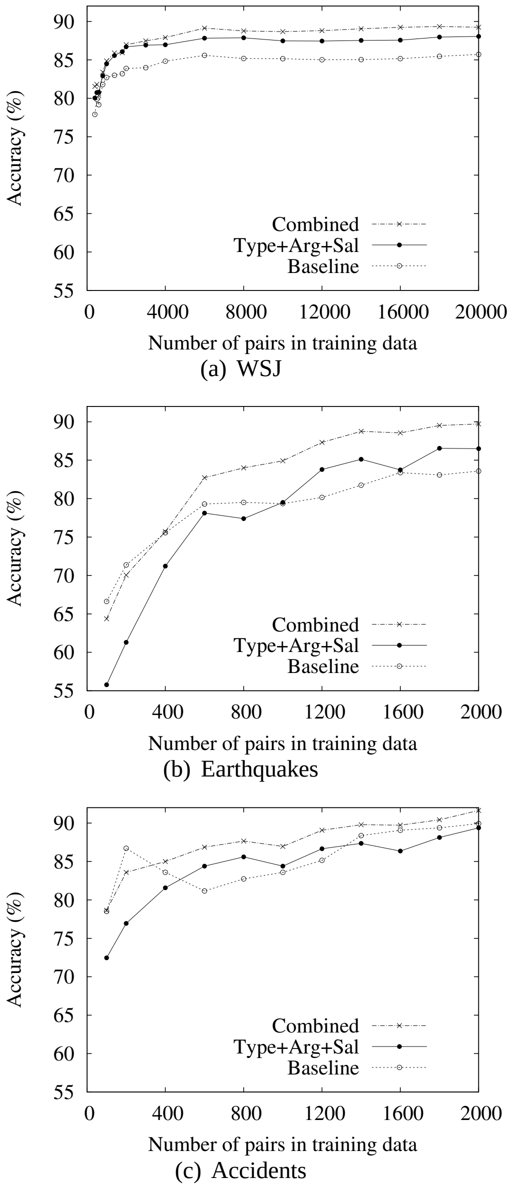
We further note that some of the permuted texts may read as coherently as the original text. This phenomenon has been observed in several natural language synthesis tasks such as generation and summarization, in which a single gold standard is inadequate to fully assess performance. As such, both automated systems and humans may actually perform better than our performance measures indicate. We leave it to future work to measure the impact of this phenomenon.

6 Analysis and Discussion

When we compare the accuracies of the full model in the three data sets (Row 2), the accuracy in the Accidents data is the highest (89.38%), followed by that in the WSJ (88.06%), with Earthquakes at the lowest (86.50%). To explain the variation, we examine the ratio between the number of the relations in the article and the article length (*i.e.*, number of sentences). This ratio is 1.22 for the Accidents source articles, 1.2 for the WSJ, and 1.08 for Earthquakes. The relation/length ratio gives us an idea of how often a sentence participates in discourse relations. A high ratio means that the article is densely interconnected by discourse relations, and may make distinguishing this article from its permutation easier compared to that for a loosely connected article.

We expect that when a text contains more discourse relation types (*i.e.*, Temporal, Contingency, Comparison, Expansion) and less EntRel and NoRel types, it is easier to compute how coherent this text is. This is because compared to EntRel and NoRel, these four discourse relations can combine to produce meaningful transitions, such as the example Text (2). To examine how this affects performance, we calculate the average ratio between the number of the four discourse relations in the permuted text and the length for the permuted text. The ratio is 0.58 for those that are correctly ranked by our system, and 0.48 for those that are incorrectly ranked, which supports our hypothesis.

We also examined the learning curves for our Type+Arg+Sal model, the baseline model, and the combined model on the data sets, as shown in Figure 2(a)–2(c). In the WSJ data, the accuracies for all three models increase rapidly as more pairs are added to the training set. After 2,000 pairs, the increase slows until 8,000 pairs, after which the curve is nearly flat. From the curves, our model consistently performs better than the baseline with a significant gap, and the combined model also consistently and significantly outperforms the other two. Only about half of the total training data is needed to reach optimal performance for all three models. The learning curves in the Earthquakes data show that the performance for all models is always increasing as more training pairs are utilized. The Type+Arg+Sal and combined models start with lower accuracies than the baseline, but catch up with it at 1,000 and 400 pairs, respectively, and consistently outperform the baseline beyond this point. On the other hand, the learning curves for the Type+Arg+Sal and baseline models in Accidents do not show any one curve consistently better than the other: our model outperforms in the middle segment but underperforms in the first and last segments. The curve for the combined model shows a consistently significant gap between it and the other two curves after the point at 400 pairs.



(a) WSJ

(b) Earthquakes

(c) Accidents

Figure 2: Learning curves for the Type+Arg+Sal, the baseline, and the combined models on the three data sets.

With the performance of the model as it is, how can future work improve upon it? We point out one weakness that we plan to explore. We use the full Type+Arg+Sal model trained on the WSJ training data to test Text (2) from the introduction. As (2) has 3 sentences, permuting it gives rise to 5 permutations. The model is able to correctly rank four of these 5 pairs. The only permutation it fails on is ($S_3 \prec S_1 \prec S_2$), when the last sentence is moved to the beginning. A very good clue of coherence in Text (2) is the explicit Comp relation between S_1 and S_2. Since this clue is retained in ($S_3 \prec S_1 \prec S_2$), it is difficult for the system to distinguish this ordering from the source. In contrast, as this clue is not present in the other four permutations, it is easier to distinguish them as incoherent. By modeling longer range discourse relation transitions, we may be able to discern these two cases.

While performance on identifying explicit discourse relations in the PDTB is as high as 93% (Pitler et al., 2008), identifying implicit ones has been shown to be a difficult task with accuracy of 40% at Level-2 types (Lin et al., 2009). As the overall performance of the PDTB parser is still less accurate than we hope it to be, we expect that our proposed model will give better performance than it does now, when the current PDTB parser performance is improved.

7 Conclusion

We have proposed a new model for discourse coherence that leverages the observation that coherent texts preferentially follow certain discourse structures. We posit that these structures can be captured in and represented by the patterns of discourse relation transitions. We first demonstrate that simply using the sequence of discourse relation transition leads to sparse features and is insufficient to distinguish coherent from incoherent text. To address this, our method transforms the discourse relation transitions into a discourse role matrix. The matrix schematically represents term occurrences in text units and associates each occurrence with its discourse roles in the text units. In our approach, n-gram sub-sequences of transitions per term in the discourse role matrix then constitute the more fine-grained evidence used in our model to distinguish coherence from incoherence.

When applied to distinguish a source text from a sentence-reordered permutation, our model significantly outperforms the previous state-of-the-art, the entity-based local coherence model. While the entity-based model captures repetitive mentions of entities, our discourse relation-based model gleans its evidence from the argumentative and discourse structure of the text. Our model is complementary to the entity-based model, as it tackles the same problem from a different perspective. Experiments validate our claim, with a combined model outperforming both single models.

The idea of modeling coherence with discourse relations and formulating it in a discourse role matrix can also be applied to other NLP tasks. We plan to apply our methodology to other tasks, such as summarization, text generation and essay scoring, which also need to produce and assess discourse coherence.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 141–148, Morristown, NJ, USA. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34:1–34, March.

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting 2004*.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Proceedings of the Conference on Human Language Technology and North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007)*, Rochester, New York, USA, April.

Robert Elwell and Jason Baldridge. 2008. Discourse connective argument identification with connective specific rankers. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2010)*, Washington, DC, USA.

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

Thorsten Joachims. 1999. Making large-scale support vector machine learning practical. In Bernhard

Schlkopf, Christopher J. C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pages 169–184. MIT Press, Cambridge, MA, USA.

Nikiforos Karamanis. 2007. Supplementing entity coherence with local rhetorical relations for information ordering. *Journal of Logic, Language and Information*, 16:445–464, October.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, UK.

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 (EMNLP 2009)*, Singapore.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A PDTB-styled end-to-end discourse parser. Technical Report TRB8/10, School of Computing, National University of Singapore, August.

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

Daniel Marcu. 1996. Distinguishing between coherent and incoherent texts. In *The Proceedings of the Student Conference on Computational Linguistics in Montreal*, pages 136–143.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48, March.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Singapore.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008) Short Papers*, Manchester, UK.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, Singapore.

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)*.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the COLING/ACL Main Conference Poster Sessions*, pages 803–810, Morristown, NJ, USA. Association for Computational Linguistics.

WenTing Wang, Jian Su, and Chew Lim Tan. 2010. Kernel based discourse relation recognition with temporal ordering information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, July.

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

Ben Wellner and James Pustejovsky. 2007. Automatically identifying the arguments of discourse connectives. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, Czech Republic.