Modeling Local Coherence: An Entity-based Approach

Regina Barzilay
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
regina@csail.mit.edu

Mirella Lapata
School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

Abstract

This paper considers the problem of automatic assessment of local coherence. We present a novel entity-based representation of discourse which is inspired by Centering Theory and can be computed automatically from raw text. We view coherence assessment as a ranking learning problem and show that the proposed discourse representation supports the effective learning of a ranking function. Our experiments demonstrate that the induced model achieves significantly higher accuracy than a state-of-the-art coherence model.

1 Introduction

A key requirement for any system that produces text is the coherence of its output. Not surprisingly, a variety of coherence theories have been developed over the years (e.g., Mann and Thomson, 1988; Grosz et al. 1995) and their principles have found application in many symbolic text generation systems (e.g., Scott and de Souza, 1990; Kibble and Power, 2004). The ability of these systems to generate high quality text, almost indistinguishable from human writing, makes the incorporation of coherence theories in robust large-scale systems particularly appealing. The task is, however, challenging considering that most previous efforts have relied on handcrafted rules, valid only for limited domains, with no guarantee of scalability or portability (Reiter and Dale, 2000). Furthermore, coherence constraints are often embedded in complex representations (e.g., Asher and Lascarides, 2003) which are hard to implement in a robust application.

This paper focuses on *local coherence*, which captures text relatedness at the level of sentence-to-sentence transitions, and is essential for generating *globally* coherent text. The key premise of our work is that the distribution of entities in locally coherent texts exhibits certain regularities. This assumption is not arbitrary — some of these regularities have been recognized in Centering Theory (Grosz et al., 1995) and other entity-based theories of discourse.

The algorithm introduced in the paper automatically abstracts a text into a set of entity transition sequences, a representation that reflects distributional, syntactic, and referential information about discourse entities. We argue that this representation of discourse allows the system to learn the properties of locally coherent texts opportunistically from a given corpus, without recourse to manual annotation or a predefined knowledge base.

We view coherence assessment as a ranking problem and present an efficiently learnable model that orders alternative renderings of the same information based on their degree of local coherence. Such a mechanism is particularly appropriate for generation and summarization systems as they can produce multiple text realizations of the same underlying content, either by varying parameter values, or by relaxing constraints that control the generation process. A system equipped with a ranking mechanism, could compare the quality of the candidate outputs, much in the same way speech recognizers employ *language models* at the sentence level.

Our evaluation results demonstrate the effectiveness of our entity-based ranking model within the general framework of coherence assessment. First, we evaluate the utility of the model in a text ordering task where our algorithm has to select a maximally coherent sentence order from a set of candidate permutations. Second, we compare the rankings produced by the model against human coherence judgments elicited for automatically generated summaries. In both experiments, our method yields

Proceedings of the 43rd Annual Meeting of the ACL, pages 141–148,
Ann Arbor, June 2005. ©2005 Association for Computational Linguistics

a significant improvement over a state-of-the-art coherence model based on Latent Semantic Analysis (Foltz et al., 1998).

In the following section, we provide an overview of existing work on the automatic assessment of local coherence. Then, we introduce our entity-based representation, and describe our ranking model. Next, we present the experimental framework and data. Evaluation results conclude the paper.

2 Related Work

Local coherence has been extensively studied within the modeling framework put forward by Centering Theory (Grosz et al., 1995; Walker et al., 1998; Strube and Hahn, 1999; Poesio et al., 2004; Kibble and Power, 2004). One of the main assumptions underlying Centering is that a text segment which foregrounds a single entity is perceived to be more coherent than a segment in which multiple entities are discussed. The theory formalizes this intuition by introducing constraints on the distribution of discourse entities in coherent text. These constraints are formulated in terms of *focus*, the most salient entity in a discourse segment, and *transition* of focus between adjacent sentences. The theory also establishes constraints on the linguistic realization of focus, suggesting that it is more likely to appear in prominent syntactic positions (such as subject or object), and to be referred to with anaphoric expressions.

A great deal of research has attempted to translate principles of Centering Theory into a robust coherence metric (Miltsakaki and Kukich, 2000; Hasler, 2004; Karamanis et al., 2004). Such a translation is challenging in several respects: one has to specify the "free parameters" of the system (Poesio et al., 2004) and to determine ways of combining the effects of various constraints. A common methodology that has emerged in this research is to develop and evaluate coherence metrics on manually annotated corpora. For instance, Miltsakaki and Kukich (2000) annotate a corpus of student essays with transition information, and show that the distribution of transitions correlates with human grades. Karamanis et al. (2004) use a similar methodology to compare coherence metrics with respect to their usefulness for text planning in generation.

The present work differs from these approaches in two key respects. First, our method does not require manual annotation of input texts. We do not aim to produce complete centering annotations; instead, our inference procedure is based on a discourse representation that preserves essential entity transition information, and can be computed automatically from raw text. Second, we learn patterns of entity distribution from a corpus, without attempting to directly implement or refine Centering constraints.

3 The Coherence Model

In this section we introduce our entity-based representation of discourse. We describe how it can be computed and how entity transition patterns can be extracted. The latter constitute a rich feature space on which probabilistic inference is performed.

Text Representation Each text is represented by an *entity grid*, a two-dimensional array that captures the distribution of discourse entities across text sentences. We follow Miltsakaki and Kukich (2000) in assuming that our unit of analysis is the traditional sentence (i.e., a main clause with accompanying subordinate and adjunct clauses). The rows of the grid correspond to sentences, while the columns correspond to discourse entities. By *discourse entity* we mean a class of coreferent noun phrases. For each occurrence of a discourse entity in the text, the corresponding grid cell contains information about its grammatical role in the given sentence. Each grid column thus corresponds to a string from a set of categories reflecting the entity's presence or absence in a sequence of sentences. Our set consists of four symbols: **s** (subject), **o** (object), **x** (neither subject nor object) and **–** (gap which signals the entity's absence from a given sentence).

Table 1 illustrates a fragment of an entity grid constructed for the text in Table 2. Since the text contains six sentences, the grid columns are of length six. Consider for instance the grid column for the entity *trial*, **[o – – – – x]**. It records that *trial* is present in sentences 1 and 6 (as **o** and **x** respectively) but is absent from the rest of the sentences.

Grid Computation The ability to identify and cluster coreferent discourse entities is an important prerequisite for computing entity grids. The same entity may appear in different linguistic forms, e.g., *Microsoft Corp.*, *Microsoft*, and *the company*, but should still be mapped to a single entry in the grid. Table 1 exemplifies the entity grid for the text in Table 2 when coreference resolution is taken into account. To automatically compute entity classes,

	Department	Trial	Microsoft	Evidence	Competitors	Markets	Products	Brands	Case	Netscape	Software	Tactics	Government	Suit	Earnings	
1	**s**	**o**	**s**	**x**	**o**	–	–	–	–	–	–	–	–	–	–	1
2	–	–	**o**	–	–	**x**	**s**	**o**	–	–	–	–	–	–	–	2
3	–	–	**s**	**o**	–	–	–	–	**s**	**o**	**o**	–	–	–	–	3
4	–	–	**s**	–	–	–	–	–	–	–	–	**s**	–	–	–	4
5	–	–	–	–	–	–	–	–	–	–	–	–	**s**	**o**	–	5
6	–	**x**	**s**	–	–	–	–	–	–	–	–	–	–	–	**o**	6

Table 1: A fragment of the entity grid. Noun phrases are represented by their head nouns.

1 [The Justice Department]$_{\mathbf{s}}$ is conducting an [anti-trust trial]$_{\mathbf{o}}$ against [Microsoft Corp.]$_{\mathbf{x}}$ with [evidence]$_{\mathbf{x}}$ that [the company]$_{\mathbf{s}}$ is increasingly attempting to crush [competitors]$_{\mathbf{o}}$.
2 [Microsoft]$_{\mathbf{o}}$ is accused of trying to forcefully buy into [markets]$_{\mathbf{x}}$ where [its own products]$_{\mathbf{s}}$ are not competitive enough to unseat [established brands]$_{\mathbf{o}}$.
3 [The case]$_{\mathbf{s}}$ revolves around [evidence]$_{\mathbf{o}}$ of [Microsoft]$_{\mathbf{s}}$ aggressively pressuring [Netscape]$_{\mathbf{o}}$ into merging [browser software]$_{\mathbf{o}}$.
4 [Microsoft]$_{\mathbf{s}}$ claims [its tactics]$_{\mathbf{s}}$ are commonplace and good economically.
5 [The government]$_{\mathbf{s}}$ may file [a civil suit]$_{\mathbf{o}}$ ruling that [conspiracy]$_{\mathbf{s}}$ to curb [competition]$_{\mathbf{o}}$ through [collusion]$_{\mathbf{x}}$ is [a violation of the Sherman Act]$_{\mathbf{o}}$.
6 [Microsoft]$_{\mathbf{s}}$ continues to show [increased earnings]$_{\mathbf{o}}$ despite [the trial]$_{\mathbf{x}}$.

Table 2: Summary augmented with syntactic annotations for grid computation.

we employ a state-of-the-art noun phrase coreference resolution system (Ng and Cardie, 2002) trained on the MUC (6–7) data sets. The system decides whether two NPs are coreferent by exploiting a wealth of features that fall broadly into four categories: lexical, grammatical, semantic and positional.

Once we have identified entity classes, the next step is to fill out grid entries with relevant syntactic information. We employ a robust statistical parser (Collins, 1997) to determine the constituent structure for each sentence, from which subjects (**s**), objects (**o**), and relations other than subject or object (**x**) are identified. Passive verbs are recognized using a small set of patterns, and the underlying deep grammatical role for arguments involved in the passive construction is entered in the grid (see the grid cell **o** for *Microsoft*, Sentence 2, Table 2).

When a noun is attested more than once with a different grammatical role in the same sentence, we default to the role with the highest grammatical ranking: subjects are ranked higher than objects, which in turn are ranked higher than the rest. For example, the entity *Microsoft* is mentioned twice in Sentence 1 with the grammatical roles **x** (for *Microsoft Corp.*) and **s** (for *the company*), but is represented only by **s** in the grid (see Tables 1 and 2).

Coherence Assessment We introduce a method for coherence assessment that is based on grid representation. A fundamental assumption underlying our approach is that the distribution of entities in coherent texts exhibits certain regularities reflected in grid topology. Some of these regularities are formalized in Centering Theory as constraints on transitions of local focus in adjacent sentences. Grids of coherent texts are likely to have some dense columns (i.e., columns with just a few gaps such as *Microsoft* in Table 1) and many sparse columns which will consist mostly of gaps (see *markets*, *earnings* in Table 1). One would further expect that entities corresponding to dense columns are more often subjects or objects. These characteristics will be less pronounced in low-coherence texts.

Inspired by Centering Theory, our analysis revolves around patterns of local entity transitions. A *local entity transition* is a sequence $\{\mathbf{s}, \mathbf{o}, \mathbf{x}, -\}^n$ that represents entity occurrences and their syntactic roles in *n* adjacent sentences. Local transitions can be easily obtained from a grid as continuous subsequences of each column. Each transition will have a certain probability in a given grid. For instance, the probability of the transition **[s –]** in the grid from Table 1 is 0.08 (computed as a ratio of its frequency (i.e., six) divided by the total number of transitions of length two (i.e., 75)). Each text can thus be viewed as a distribution defined over transition types. We believe that considering all entity transitions may uncover new patterns relevant for coherence assessment.

We further refine our analysis by taking into account the salience of discourse entities. Centering and other discourse theories conjecture that the way an entity is introduced and mentioned depends on its global role in a given discourse. Therefore, we discriminate between transitions of salient entities and the rest, collecting statistics for each group separately. We identify salient entities based on their

	S S	S O	S X	S –	O S	O O	O X	O –	X S	X O	X X	X –	– S	– O	– X	– –
d_1	0	0	0	.03	0	0	0	.02	.07	0	0	.12	.02	.02	.05	.25
d_2	0	0	0	.02	0	.07	0	.02	0	0	.06	.04	0	0	0	.36
d_3	.02	0	0	.03	0	0	0	.06	0	0	0	.05	.03	.07	.07	.29

Table 3: Example of a feature-vector document representation using all transitions of length two given syntactic categories: **S**, **O**, **X**, and **–**.

frequency,[1] following the widely accepted view that the occurrence frequency of an entity correlates with its discourse prominence (Morris and Hirst, 1991; Grosz et al., 1995).

Ranking We view coherence assessment as a ranking learning problem. The ranker takes as input a set of alternative renderings of the same document and ranks them based on their degree of local coherence. Examples of such renderings include a set of different sentence orderings of the same text and a set of summaries produced by different systems for the same document. Ranking is more suitable than classification for our purposes since in text generation, a system needs a scoring function to compare among alternative renderings. Furthermore, it is clear that coherence assessment is not a categorical decision but a graded one: there is often no single coherent rendering of a given text but many different possibilities that can be partially ordered.

As explained previously, coherence constraints are modeled in the grid representation implicitly by entity transition sequences. To employ a machine learning algorithm to our problem, we encode transition sequences explicitly using a standard feature vector notation. Each grid rendering j of a document d_i is represented by a feature vector $\Phi(x_{ij}) = (p_1(x_{ij}), p_2(x_{ij}), \ldots, p_m(x_{ij}))$, where m is the number of all predefined entity transitions, and $p_t(x_{ij})$ the probability of transition t in grid x_{ij}. Note that considerable latitude is available when specifying the transition types to be included in a feature vector. These can be all transitions of a given length (e.g., two or three) or the most frequent transitions within a document collection. An example of a feature space with transitions of length two is illustrated in Table 3.

The training set consists of ordered pairs of renderings (x_{ij}, x_{ik}), where x_{ij} and x_{ik} are renderings of the same document d_i, and x_{ij} exhibits a higher degree of coherence than x_{ik}. Without loss of generality, we assume $j > k$. The goal of the training procedure is to find a parameter vector $\vec{w}$ that yields a "ranking score" function $\vec{w} \cdot \Phi(x_{ij})$, which minimizes the number of violations of pairwise rankings provided in the training set. Thus, the ideal $\vec{w}$ would satisfy the condition $\vec{w} \cdot (\Phi(x_{ij}) - \Phi(x_{ik})) > 0 \; \forall j,i,k$ such that $j > k$. The problem is typically treated as a Support Vector Machine constraint optimization problem, and can be solved using the search technique described in Joachims (2002a). This approach has been shown to be highly effective in various tasks ranging from collaborative filtering (Joachims, 2002a) to parsing (Toutanova et al., 2004).

In our ranking experiments, we use Joachims' (2002a) SVM^{light} package for training and testing with all parameters set to their default values.

4 Evaluation Set-Up

In this section we describe two evaluation tasks that assess the merits of the coherence modeling framework introduced above. We also give details regarding our data collection, and parameter estimation. Finally, we introduce the baseline method used for comparison with our approach.

4.1 Text Ordering

Text structuring algorithms (Lapata, 2003; Barzilay and Lee, 2004; Karamanis et al., 2004) are commonly evaluated by their performance at information-ordering. The task concerns determining a sequence in which to present a pre-selected set of information-bearing items; this is an essential step in concept-to-text generation, multi-document summarization, and other text-synthesis problems. Since local coherence is a key property of any well-formed text, our model can be used to rank alternative sentence orderings. We do not assume that local coherence is sufficient to uniquely determine the best ordering — other constraints clearly play a role here. However, we expect that the accuracy of a coherence model is reflected in its performance in the ordering task.

Data To acquire a large collection for training and testing, we create synthetic data, wherein the candidate set consists of a source document and permutations of its sentences. This framework for data acquisition is widely used in evaluation of ordering algorithms as it enables large scale automatic evalu-

[1] The frequency threshold is empirically determined on the development set. See Section 5 for further discussion.

ation. The underlying assumption is that the original sentence order in the source document must be coherent, and so we should prefer models that rank it higher than other permutations. Since we do not know the relative quality of different permutations, our corpus includes only pairwise rankings that comprise the original document and one of its permutations. Given k original documents, each with n randomly generated permutations, we obtain $k \cdot n$ (trivially) annotated pairwise rankings for training and testing.

Using the technique described above, we collected data in two different genres: newspaper articles and accident reports written by government officials. The first collection consists of Associated Press articles from the North American News Corpus on the topic of natural disasters. The second includes narratives from the National Transportation Safety Board's database[2]. Both sets have documents of comparable length – the average number of sentences is 10.4 and 11.5, respectively. For each set, we used 100 source articles with 20 randomly generated permutations for training. The same number of pairwise rankings (i.e., 2000) was used for testing. We held out 10 documents (i.e., 200 pairwise rankings) from the training data for development purposes.

4.2 Summary Evaluation

We further test the ability of our method to assess coherence by comparing model induced rankings against rankings elicited by human judges. Admittedly, the information ordering task only partially approximates degrees of coherence violation using different sentence permutations of a source document. A stricter evaluation exercise concerns the assessment of texts with naturally occurring coherence violations as perceived by human readers. A representative example of such texts are automatically generated summaries which often contain sentences taken out of context and thus display problems with respect to local coherence (e.g., dangling anaphors, thematically unrelated sentences). A model that exhibits high agreement with human judges not only accurately captures the coherence properties of the summaries in question, but ultimately holds promise for the automatic evaluation of machine-generated texts. Existing automatic evaluation measures such as BLEU (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2003), are not designed for the coherence assessment task, since they focus on content similarity between system output and reference texts.

Data Our evaluation was based on materials from the Document Understanding Conference (DUC, 2003), which include multi-document summaries produced by human writers and by automatic summarization systems. In order to learn a ranking, we require a set of summaries, each of which have been rated in terms of coherence. We therefore elicited judgments from human subjects.[3] We randomly selected 16 input document clusters and five systems that had produced summaries for these sets, along with summaries composed by several humans. To ensure that we do not tune a model to a particular system, we used the output summaries of distinct systems for training and testing. Our set of training materials contained $4 \cdot 16$ summaries (average length 4.8), yielding $\binom{4}{2} \cdot 16 = 96$ pairwise rankings. In a similar fashion, we obtained 32 pairwise rankings for the test set. Six documents from the training data were used as a development set.

Coherence ratings were obtained during an elicitation study by 177 unpaid volunteers, all native speakers of English. The study was conducted remotely over the Internet. Participants first saw a set of instructions that explained the task, and defined the notion of coherence using multiple examples. The summaries were randomized in lists following a Latin square design ensuring that no two summaries in a given list were generated from the same document cluster. Participants were asked to use a seven point scale to rate how coherent the summaries were without having seen the source texts. The ratings (approximately 23 per summary) given by our subjects were averaged to provide a rating between 1 and 7 for each summary.

The reliability of the collected judgments is crucial for our analysis; we therefore performed several tests to validate the quality of the annotations. First, we measured how well humans agree in their coherence assessment. We employed leave-one-out resampling[4] (Weiss and Kulikowski, 1991), by correlating the data obtained from each participant with the mean coherence ratings obtained from all other participants. The inter-subject agree-

[2] The collections are available from `http://www.csail.mit.edu/regina/coherence/`.

[3] The ratings are available from `http://homepages.inf.ed.ac.uk/mlap/coherence/`.

[4] We cannot apply the commonly used Kappa statistic for measuring agreement since it is appropriate for nominal scales, whereas our summaries are rated on an ordinal scale.

ment was $r = .768$. Second, we examined the effect of different types of summaries (human- vs. machine-generated.) An ANOVA revealed a reliable effect of summary type: $F(1;15) = 20.38$, $p < 0.01$ indicating that human summaries are perceived as significantly more coherent than system-generated ones. Finally, the judgments of our participants exhibit a significant correlation with DUC evaluations ($r = .41$, $p < 0.01$).

4.3 Parameter Estimation

Our model has two free parameters: the frequency threshold used to identify salient entities and the length of the transition sequence. These parameters were tuned separately for each data set on the corresponding held-out development set. For our ordering and summarization experiments, optimal salience-based models were obtained for entities with frequency ≥ 2. The optimal transition length was ≤ 3 for ordering and ≤ 2 for summarization.

4.4 Baseline

We compare our algorithm against the coherence model proposed by Foltz et al. (1998) which measures coherence as a function of semantic relatedness between adjacent sentences. Semantic relatedness is computed automatically using Latent Semantic Analysis (LSA, Landauer and Dumais 1997) from raw text without employing syntactic or other annotations. This model is a good point of comparison for several reasons: (a) it is fully automatic, (b) it is a not a straw-man baseline; it correlates reliably with human judgments and has been used to analyze discourse structure, and (c) it models an aspect of coherence which is orthogonal to ours (their model is lexicalized).

Following Foltz et al. (1998) we constructed vector-based representations for individual words from a lemmatized version of the North American News Text Corpus[5] (350 million words) using a term-document matrix. We used singular value decomposition to reduce the semantic space to 100 dimensions obtaining thus a space similar to LSA. We represented the meaning of a sentence as a vector by taking the mean of the vectors of its words. The similarity between two sentences was determined by measuring the cosine of their means. An overall text coherence measure was obtained by averaging the cosines for all pairs of adjacent sentences.

In sum, each text was represented by a single feature, its sentence-to-sentence semantic similarity. During training, the ranker learns an appropriate threshold value for this feature.

4.5 Evaluation Metric

Model performance was assessed in the same way for information ordering and summary evaluation. Given a set of pairwise rankings, we measure accuracy as the ratio of correct predictions made by the model over the size of the test set. In this setup, random prediction results in an accuracy of 50%.

5 Results

The evaluation of our coherence model was driven by two questions: (1) How does the proposed model compare to existing methods for coherence assessment that make use of distinct representations? (2) What is the contribution of linguistic knowledge to the model's performance? Table 4 summarizes the accuracy of various configurations of our model for the ordering and coherence assessment tasks.

We first compared a linguistically rich grid model that incorporates coreference resolution, expressive syntactic information, and a salience-based feature space (Coreference+Syntax+Salience) against the LSA baseline (LSA). As can be seen in Table 4, the grid model outperforms the baseline in both ordering and summary evaluation tasks, by a wide margin. We conjecture that this difference in performance stems from the ability of our model to discriminate between various patterns of local sentence transitions. In contrast, the baseline model only measures the degree of overlap across successive sentences, without taking into account the properties of the entities that contribute to the overlap. Not surprisingly, the difference between the two methods is more pronounced for the second task — summary evaluation. Manual inspection of our summary corpus revealed that low-quality summaries often contain repetitive information. In such cases, simply knowing about high cross-sentential overlap is not sufficient to distinguish a repetitive summary from a well-formed one.

In order to investigate the contribution of linguistic knowledge on model performance we compared the full model introduced above against models using more impoverished representations. We focused on three sources of linguistic knowledge — syntax, coreference resolution, and salience — which play

[5] Our selection of this corpus was motivated by its similarity to the DUC corpus which primarily consists of news stories.

Model	Ordering (Set1)	Ordering (Set2)	Summarization
Coreference+Syntax+Salience	**87.3**	**90.4**	**68.8**
Coreference+Salience	86.9	88.3	62.5
Syntax+Salience	83.4	89.7	81.3
Coreference+Syntax	76.5	88.8	75.0
LSA	72.1	72.1	25.0

Table 4: Ranking accuracy measured as the fraction of correct pairwise rankings in the test set.

a prominent role in Centering analyses of discourse coherence. An additional motivation for our study is exploration of the trade-off between robustness and richness of linguistic annotations. NLP tools are typically trained on human-authored texts, and may deteriorate in performance when applied to automatically generated texts with coherence violations.

Syntax To evaluate the effect of syntactic knowledge, we eliminated the identification of grammatical relations from our grid computation and recorded solely whether an entity is present or absent in a sentence. This leaves only the coreference and salience information in the model, and the results are shown in Table 4 under (Coreference+Salience). The omission of syntactic information causes a uniform drop in performance on both tasks, which confirms its importance for coherence analysis.

Coreference To measure the effect of fully-fledged coreference resolution, we constructed entity classes simply by clustering nouns on the basis of their identity. In other words, each noun in a text corresponds to a different entity in a grid, and two nouns are considered coreferent only if they are identical. The performance of the model (Syntax+Salience) is shown in the third row of Table 4.

While coreference resolution improved model performance in ordering, it caused a decrease in accuracy in summary evaluation. This drop in performance can be attributed to two factors related to the nature of our corpus — machine-generated texts. First, an automatic coreference resolution tool expectedly decreases in accuracy because it was trained on well-formed human-authored texts. Second, automatic summarization systems do not use anaphoric expressions as often as humans do. Therefore, a simple entity clustering method is more suitable for automatic summaries.

Salience Finally, we evaluate the contribution of salience information by comparing our original model (Coreference+Syntax+Salience) which accounts separately for patterns of salient and non-salient entities against a model that does not attempt to discriminate between them (Coreference+Syntax). Our results on the ordering task indicate that models that take salience information into account consistently outperform models that do not. The effect of salience is less pronounced for the summarization task when it is combined with coreference information (Coreference + Salience). This is expected, since accurate identification of coreferring entities is prerequisite to deriving accurate salience models. However, as explained above, our automatic coreference tool introduces substantial noise in our representation. Once this noise is removed (see Syntax+Salience), the salience model has a clear advantage over the other models.

6 Discussion and Conclusions

In this paper we proposed a novel framework for representing and measuring text coherence. Central to this framework is the *entity grid* representation of discourse which we argue captures important patterns of sentence transitions. We re-conceptualize coherence assessment as a ranking task and show that our entity-based representation is well suited for learning an appropriate ranking function; we achieve good performance on text ordering and summary coherence evaluation.

On the linguistic side, our results yield empirical support to some of Centering Theory's main claims. We show that coherent texts are characterized by transitions with particular properties which do not hold for all discourses. Our work, however, not only validates these findings, but also quantitatively measures the predictive power of various linguistic features for the task of coherence assessment.

An important future direction lies in augmenting our entity-based model with lexico-semantic knowledge. One way to achieve this goal is to cluster entities based on their semantic relatedness, thereby creating a grid representation over lexical chains (Morris and Hirst, 1991). An entirely different approach is to develop fully lexicalized models, akin to traditional language models. Cache language models (Kuhn and Mori, 1990) seem particularly promising in this context.

In the discourse literature, entity-based theories are primarily applied at the level of local coherence, while relational models, such as Rhetorical Structure Theory (Mann and Thomson, 1988; Marcu, 2000), are used to model the global structure of discourse. We plan to investigate how to combine the two for improved prediction on both local and global levels, with the ultimate goal of handling longer texts.

Acknowledgments

The authors acknowledge the support of the National Science Foundation (Barzilay; CAREER grant IIS-0448168 and grant IIS-0415865) and EPSRC (Lapata; grant GR/T04540/01). We are grateful to Claire Cardie and Vincent Ng for providing us the results of their system on our data. Thanks to Eli Barzilay, Eugene Charniak, Michael Elhadad, Noemie Elhadad, Frank Keller, Alex Lascarides, Igor Malioutov, Smaranda Muresan, Martin Rinard, Kevin Simler, Caroline Sporleder, Chao Wang, Bonnie Webber and three anonymous reviewers for helpful comments and suggestions. Any opinions, findings, and conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the National Science Foundation or EPSRC.

References

N. Asher, A. Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

R. Barzilay, L. Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL*, 113–120.

M. Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the ACL/EACL*, 16–23.

P. W. Foltz, W. Kintsch, T. K. Landauer. 1998. Textual coherence using latent semantic analysis. *Discourse Processes*, 25(2&3):285–307.

B. Grosz, A. K. Joshi, S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

L. Hasler. 2004. An investigation into the use of centering transitions for summarisation. In *Proceedings of the 7th Annual CLUK Research Colloquium*, 100–107, University of Birmingham.

T. Joachims. 2002a. Optimizing search engines using clickthrough data. In *Proceesings of KDD*, 133–142.

N. Karamanis, M. Poesio, C. Mellish, J. Oberlander. 2004. Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. In *Proceedings of the ACL*, 391–398.

R. Kibble, R. Power. 2004. Optimising referential coherence in text generation. *Computational Linguistics*, 30(4):401–416.

R. Kuhn, R. D. Mori. 1990. A cache-based natural language model for speech recognition. *IEEE Transactions on PAMI*, 12(6):570–583.

T. K. Landauer, S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

M. Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the ACL*, 545–552.

C.-Y. Lin, E. H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL*, 71–78.

W. C. Mann, S. A. Thomson. 1988. Rhetorical structure theory. *Text*, 8(3):243–281.

D. Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

E. Miltsakaki, K. Kukich. 2000. The role of centering theory's rough-shift in the teaching and evaluation of writing skills. In *Proceedings of the ACL*, 408–415.

J. Morris, G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 1(17):21–43.

V. Ng, C. Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the ACL*, 104–111.

K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, 311–318.

M. Poesio, R. Stevenson, B. D. Eugenio, J. Hitzeman. 2004. Centering: a parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

E. Reiter, R. Dale. 2000. *Building Natural-Language Generation Systems*. Cambridge University Press.

D. Scott, C. S. de Souza. 1990. Getting the message across in RST-based text generation. In R. Dale, C. Mellish, M. Zock, eds., *Current Research in Natural Language Generation*, 47–73. Academic Press.

M. Strube, U. Hahn. 1999. Functional centering – grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

K. Toutanova, P. Markova, C. D. Manning. 2004. The leaf projection path view of parse trees: Exploring string kernels for HPSG parse selection. In *Proceedings of the EMNLP*, 166–173.

M. Walker, A. Joshi, E. Prince, eds. 1998. *Centering Theory in Discourse*. Clarendon Press.

S. M. Weiss, C. A. Kulikowski. 1991. *Computer Systems that Learn: Classification and Prediction Methods from, Statistics, Neural Nets, Machine Learning, and Expert Systems*. Morgan Kaufmann.