Automatic Evaluation of Text Coherence: Models and Representations

Mirella Lapata
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
mlap@inf.ed.ac.uk

Regina Barzilay
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
G-468, the Stata Center
Cambridge, MA 14853
regina@csail.mit.edu

Abstract

This paper investigates the automatic evaluation of text coherence for machine-generated texts. We introduce a fully-automatic, linguistically rich model of local coherence that correlates with human judgments. Our modeling approach relies on shallow text properties and is relatively inexpensive. We present experimental results that assess the predictive power of various discourse representations proposed in the linguistic literature. Our results demonstrate that certain models capture complementary aspects of coherence and thus can be combined to improve performance.

1 Introduction

The use of automatic methods for evaluating machine-generated text is quickly becoming mainstream in machine translation, text generation, and summarization. These methods target various dimensions of text quality, ranging from sentence grammaticality to content selection [Bangalore *et al.*, 2000; Papineni *et al.*, 2002; Hovy and Lin, 2003]. However, a fundamental aspect of text structure — its *coherence* — is still evaluated manually.

Coherence is a property of well-written texts that makes them easier to read and understand than a sequence of randomly strung sentences. Although the same information can be organized in multiple ways to create a coherent text, some forms of text organization will be indisputably judged incoherent. The automatically-generated summary in Table 1 is an example of a badly organized text: the presence of thematically unrelated sentences and non-canonical time sequences makes it nearly impossible to comprehend. We want to develop an automatic method that can distinguish such a summary from a text with no coherence violations.

This paper focuses on *local coherence*, which captures text organization at the level of sentence to sentence transitions, and is undoubtedly necessary for global coherence. This topic has received much attention in the linguistic literature [Grosz *et al.*, 1995; Morris and Hirst, 1991; Halliday and Hasan, 1976] and a variety of qualitative models have been proposed. The emphasis of our work is on *quantitative* models of local coherence that can be efficiently computed from raw text, and readily utilized for automatic evaluation. Our approach relies on shallow text properties that can be easily identified and quantified without recourse to elaborate world knowledge and handcrafted rules.

Newspapers reported Wednesday that three top Libyan officials have been tried and jailed in the Lockerbie case. Secretary-General Kofi Annan said Wednesday that he may travel to Libya next week in hopes of closing a deal. The sanctions were imposed to force Libyan leader Moammar Gadhafi to turn the men over. Louis Farrakhan, the leader of a U.S. Muslim group, met with Gadhafi and congratulated him on his recovery from a hip injury.

Table 1: A low-coherence summary

The coherence models described in this paper fall into two broad classes that capture orthogonal dimensions of entity distribution in discourse. The first class incorporates *syntactic* aspects of text coherence and characterizes how mentions of the same entity in different syntactic positions are spread across adjacent sentences. Inspired by Centering Theory [Grosz *et al.*, 1995], our algorithm abstracts a collection of coherent texts into a set of entity transition sequences, and defines a probabilistic model over their distribution. Given a new text, the model evaluates its coherence by computing the probability of its entity transitions according to the training data.

The second, *semantic*, class of models quantifies local coherence as the degree of connectivity across text sentences. This approach is motivated by the findings of Halliday and Hasan [1976] who emphasized the role of lexical cohesion in text coherence. The key parameter of these models is the definition of semantic relatedness (as a proxy for connectivity) among text entities. We explore a wealth of similarity measures, ranging from distributional to taxonomy-based, to find an optimal lexico-semantic representation for the coherence assessment task.

We employ our models to evaluate the coherence of multidocument summaries produced by systems that participated in the Document Understanding Conference. We acquire coherence ratings for a large collection of machine-generated summaries by eliciting judgments from human subjects, and examine the extent to which the predictions of various models correlate with human intuitions. Our experiments demonstrate that while several models exhibit a statistically significant agreement with human ratings, a model that fuses the syntactic and the semantic views yields substantial improvement over any single model. Our contributions are twofold:

Modeling We present a fully automatic, linguistically rich model of local coherence that correlates with human judgments. This model is the first attempt, to our knowledge, to automatically evaluate the local coherence of machine-generated texts.

Linguistic Analysis We present experimental results that assess the predictive power of various knowledge sources and discourse representations proposed in the linguistic literature. In particular, we show that certain models capture complementary aspects of coherence and thus can be combined to improve performance.

2 Related Work

Most of the work on automatic coherence evaluation has been done in the context of automatic essay scoring. Higgins *et al.* [2004] develop a system that assesses global aspects of coherence in student essays. They use a manually annotated corpus of essays to learn which types of discourse segments can cause breakdowns in coherence. Other approaches focus on local coherence. Miltsakaki and Kukich [2004] manually annotate a corpus of student essays with entity transition information, and show that the distribution of transition types correlates with human grades. Foltz *et al.* [1998] propose a model of local coherence that presupposes no manual coding. A text is considered coherent if it exhibits a high degree of meaning overlap between adjacent sentences. They employ a vector-based representation of lexical meaning and assess semantic relatedness by measuring the distance between sentence pairs. Foltz *et al.* report that the model correlates reliably with human judgments and can be used to analyze discourse structure. The success of this approach motivates our research on semantic association models of coherence.

Previous work has primarily focused on human authored texts and has typically utilized a single source of information for coherence assesment. In contrast, we concentrate on machine generated texts and assess which knowledge sources are appropriate for measuring local coherence. We introduce novel models but also assess whether previously proposed ones (e.g., Foltz *et al.* [1998]) generalize to automatically generated texts. As a byproduct of our main investigation, we also examine whether humans can reliably rate texts in terms of coherence, thus undertaking a large-scale judgment elicitation study.

3 Models of Local Coherence

In this section we introduce two classes of coherence models, and describe how they can be used in automatic evaluation. We motivate their construction, present a corresponding discourse representation, and an inference procedure.

3.1 The Syntactic View

Linguistic Motivation Centering theory [Grosz *et al.*, 1995; Walker *et al.*, 1998] is one of the most influential frameworks for modeling local coherence. Fundamental in Centering's study of discourse is the way entities are introduced and discussed. The theory asserts that discourse segments in which successive utterances mention the same entities are *more coherent* than discourse segments in which multiple entities are discussed. Coherence analysis revolves around *patterns of local entity transitions* which specify how the focus of discourse changes from sentence to sentence. The key assumption is that certain types of entity transitions are likely to appear in locally coherent discourse. Centering also establishes constraints on the linguistic realization of focus, suggesting that focused entities are likely to occupy prominent syntactic positions such as subject or object.

1. [Former Chilean dictator Augusto Pinochet]$_{\mathbf{O}}$, was arrested in [London]$_{\mathbf{X}}$ on [14 October]$_{\mathbf{X}}$ 1998.
2. [Pinochet]$_{\mathbf{S}}$, 82, was recovering from [surgery]$_{\mathbf{X}}$.
3. [The arrest]$_{\mathbf{S}}$ was in [response]$_{\mathbf{X}}$ to [an extradition warrant]$_{\mathbf{X}}$ served by [a Spanish judge]$_{\mathbf{S}}$.
4. [Pinochet]$_{\mathbf{O}}$ was charged with murdering [thousands]$_{\mathbf{O}}$, including many [Spaniards]$_{\mathbf{O}}$.
5. [Pinochet]$_{\mathbf{S}}$ is awaiting [a hearing]$_{\mathbf{O}}$, [his fate]$_{\mathbf{X}}$ in [the balance]$_{\mathbf{X}}$.
6. [American scholars]$_{\mathbf{S}}$ applauded the [arrest]$_{\mathbf{O}}$.

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

	Dictator	Augusto	Pinochet	London	October	Surgery	Arrest	Extradition	Warrant	Judge	Thousands	Spaniards	Hearing	Fate	Balance	Scholars	
1	**O**	**O**	**O**	**X**	**X**	–	–	–	–	–	–	–	–	–	–	–	1
2	–	–	**S**	–	–	**X**	–	–	–	–	–	–	–	–	–	–	2
3	–	–	–	–	–	–	**S**	**X**	**X**	**S**	–	–	–	–	–	–	3
4	–	–	**O**	–	–	–	–	–	–	–	**O**	**O**	–	–	–	–	4
5	–	–	**S**	–	–	–	–	–	–	–	–	–	**O**	**X**	**X**	–	5
6	–	–	–	–	–	–	**O**	–	–	–	–	–	–	–	–	**S**	6

Table 3: An entity grid

Discourse representation To expose entity transition patterns characteristic of coherent texts, we represent a text by an *entity grid*. The grid's columns correspond to discourse entities, while the rows correspond to utterances (see Table 3). We follow Miltsakaki and Kukich [2004] in assuming that an utterance is a traditional sentence (i.e., a main clause with accompanying subordinate and adjunct clauses).

Grid columns record an entity's presence or absence in a sequence of sentences $(S_1, \ldots, S_n)$. More specifically, each grid cell represents the role r_{ij} of entity e_j in a given sentence S_i. Grammatical roles reflect whether an entity is a subject (**S**), object (**O**), neither (**X**) or simply absent (**–**). Table 3 illustrates an entity grid constructed for the text in Table 2. Since the text contains six sentences, the grid columns are of length six. As an example consider the grid column for the entity *arrest*, **[– – S – – O]**. It records that *arrest* is present in sentence 3 as a subject and in sentence 6 as an object, but is absent from the rest of the sentences.

Ideally, each entity in the grid should represent an equivalence class of coreferent nouns (e.g., *Former Chilean dictator Augusto Pinochet* and *Pinochet* refer to the same entity). However, the automatic construction of such an entity grid requires a robust coreference tool, able to accurately process texts with coherence violations. Since coreference tools are typically trained on coherent texts, this requirement is hard to satisfy. Instead, we employ a simplified representation: each noun in a text corresponds to a different entity in the grid. The

simplification allows us to capture noun-coreference, albeit in a shallow manner — only exact repetitions are considered coreferent. In practice, this means that named entities and compound nouns will be treated as denoting more than one entity. For instance, the NP *Former Chilean dictator Augusto Pinochet* will be mapped to three entities: *dictator*, *Augusto*, and *Pinochet*. We further assume that each noun within an NP bears the same grammatical role as the NP head. Thus, all three nouns in the above NP will be labeled as objects. 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 (i.e., **s** > **o** > **x**).

Entity grids can be straightforwardly computed provided that an accurate parser is available. In the experiments reported throughout this paper we employed Collins' [1998] state-of-the-art statistical parser to identify discourse entities and their grammatical roles. Entities involved in passive constructions were identified using a small set of patterns and their corresponding deep grammatical role was entered in the grid (see the grid cell **o** for *Pinochet*, Sentence 1, Table 3).

Inference A fundamental assumption underlying our inference mechanism is that the distribution of entities in coherent texts exhibit certain regularities reflected in the topology of grid columns. Grids of coherent texts are likely to have some dense columns (i.e., columns with just a few gaps such as *Pinochet* in Table 3) and many sparse columns which will consist mostly of gaps (see *London, judge* in Table 3). 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.

We define the coherence of a text T $(S_1,\ldots,S_n)$ with entities $e_1 \ldots e_m$ as a joint probability distribution that governs how entities are distributed across document sentences:

$$P_{coherence}(T) = P(e_1 \ldots e_m; S_1 \ldots S_n) \quad (1)$$

We further assume (somewhat simplistically) that an entity is selected into a document independently of other entities:

$$P_{coherence}(T) \approx \prod_{j=1}^{m} P(e_j; S_1 \ldots S_n) \quad (2)$$

To give a concrete example, we will rate the coherence of the text in Table 3 by multiplying together the probabilities of the entities *Dictator, Augusto, Pinochet*, etc.

We define $P(e_j; S_1 \ldots S_n)$ as a probability distribution over transition sequences for entity e_j across all n sentences of text T. $P(e_j; S_1 \ldots S_n)$ is estimated from grid columns, as observed in a corpus of coherent texts:

$$\begin{aligned} P(e_j; S_1 \ldots S_n) &= P(r_{1,j} \ldots r_{n,j}) \\ &= \prod_{i=1}^{n} P(r_{i,j} | r_{1,j} \ldots r_{(i-1),j}) \\ &\approx \prod_{i=1}^{n} P(r_{i,j} | r_{(i-h),j} \ldots r_{(i-1),j}) \end{aligned} \quad (3)$$

where $r_{i,j}$ represents the grammatical role for entity e_j in sentence i (see the grid columns in Table 3). The estimates for $P(r_{i,j}|r_{1,j} \ldots r_{(i-1),j})$ are obtained using the standard Markov assumption of independence, where h is the history size. Assuming a first-order Markov model, $P(Pinochet; S_1 \ldots S_6)$ will be estimated by multiplying together $P(\mathbf{o})$, $P(\mathbf{s}|\mathbf{o})$, $P(-|\mathbf{s})$, $P(\mathbf{o}|-)$, $P(\mathbf{s}|\mathbf{o})$, and $P(-|\mathbf{s})$ (see the column for *Pinochet* in Table 3).

To compare texts with variable lengths and entities, the probabilities for individual columns are normalized by column length (n) and the probability of the entire text is normalized by the number of columns (m):

$$P_{coherence}(T) \approx \frac{1}{m \cdot n} \sum_{j=1}^{m} \sum_{i=1}^{n} \log P(r_{i,j} | r_{(i-h),j} \ldots r_{(i-1),j}) \quad (4)$$

Once the model is trained on a corpus of coherent texts, it can be used to assess the coherence of unseen texts. Texts that are given a high probability $P_{coherence}(T)$ will be deemed more coherent than those given low probability.

Notice that the model is unlexicalized, i.e., the estimation of $P(e_j; S_1 \ldots S_n)$ is not entity-specific. In practice, this means that entities with the same column topology (e.g., *London*, *October* in Table 3) will be given the same probability.

3.2 The Semantic View

Linguistic motivation An important factor in text comprehension is the degree to which sentences and phrases are linked together. The observation dates back to Halliday and Hasan [1976] who stressed the role of lexical cohesion in text coherence. A number of linguistic devices — entity repetition, synonymy, hyponymy, and meronymy — are considered to contribute to the "continuity of lexical meaning" observed in coherent text. Morris and Hirst [1991] represent lexical cohesion via *lexical chains*, i.e., sequences of related words spanning a topical text unit, thus formalizing the intuition that coherent units will have a high concentration of dense chains. They argue that the distribution of lexical chains is a surface indicator of the structure of coherent discourse.

Discourse representation The key premise behind lexical chains is that coherent texts will contain a high number of semantically related words. This allows for a particularly simple representation of discourse that does not take account of syntactic structure or even word order within a sentence. Each sentence is thus represented as a bag of words. These can be all words in the document (excluding function words) or selected grammatical categories (e.g., verbs or nouns). For our models, we assume that each sentence is represented by a set of nouns.

Central to this representation is the ability to measure semantic similarity. Different coherence models can be therefore defined according to the chosen similarity measure.

Inference Measuring local coherence amounts to quantifying the degree of semantic relatedness between sentences. Specifically, the coherence of text T is measured by taking the mean of all individual transitions.

$$coherence(T) = \frac{\sum_{i=1}^{n-1} sim(S_i, S_{i+1})}{n-1} \quad (5)$$

where $sim(S_i, S_{i+1})$ is a measure of similarity between sentences S_i and S_{i+1}.

We have experimented with three broad classes of models which employ *word-based*, *distributional*, and *taxonomy-based similarity measures*.

In its simplest form, semantic similarity can be operationalized in terms of word overlap:

$$sim(S_1, S_2) = \frac{2|words(S_1) \cap words(S_2)|}{(|words(S_1)| + |words(S_2)|)} \quad (6)$$

where $words(S_i)$ is the set of words in sentence i. The main drawback of this measure is that it will indicate low-coherence for sentence pairs that have no words in common, even though they may be semantically related.

Measures of distributional similarity, however, go beyond mere repetition. Words are considered similar if they occur within similar contexts. The semantic properties of words are captured in a multi-dimensional space by vectors that are constructed from large bodies of text by observing the distributional patterns of co-occurrence with their neighboring words. For modeling coherence, we want to be able to compare the similarity of sentences rather than words (see (4)). Our computation of sentence similarity follows the method described in Foltz [1998]: each sentence is represented by the mean (centroid) of the vectors of its words, and the similarity between two sentences is determined by the cosine of their means.

$$\begin{aligned} sim(S_1,S_2) &= \cos(\mu(\vec{S_1}),\mu(\vec{S_2})) \\ &= \frac{\sum_{j=1}^{n} \mu_j(\vec{S_1})\mu_j(\vec{S_2})}{\sqrt{\sum_{j=1}^{n}(\mu_j(\vec{S_1}))^2}\sqrt{\sum_{j=1}^{n}(\mu_j(\vec{S_2}))^2}} \end{aligned} \quad (7)$$

where $\mu(\vec{S_i}) = \frac{1}{|S_i|}\sum_{\vec{w}\in S_i}\vec{w}$, and $\vec{w}$ is the vector for word w.

An alternative to inducing word similarity relationships from co-occurrence statistics in a corpus, is to employ a manually crafted resource such as WordNet [Fellbaum, 1998]. WordNet-based similarity measures have been shown to correlate reliably with human similarity judgments and have been used in a variety of applications ranging from the detection of malapropisms to word sense disambiguation (see Budanitsky and Hirst [2001] for an extensive survey). We employed five measures commonly cited in the literature. Two of these measures [Hirst and St-Onge, 1998; Lesk, 1986] define similarity solely in terms of the taxonomy, whereas the other three are based on information theory [Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998] and combine taxonomic information with corpus counts. By considering a broad range of these measures, we should be able to distill which ones are best suited for coherence assessment.

All WordNet-based measures compute a similarity score at the sense level. We define the similarity of two sentences S_1 and S_2 as:

$$sim(S_1,S_2) = \frac{\sum_{\substack{w_1\in S_1 \\ w_2\in S_2}} \operatorname{argmax}_{\substack{c_1\in senses(w_1) \\ c_2\in senses(w_2)}} sim(c_1,c_2)}{|S_1||S_2|} \quad (8)$$

where $|S_i|$ is the number of words in sentence i. Since the appropriate senses for words w_1 and w_2 are not known, our measure selects the senses which maximize $sim(c_1,c_2)$.

4 Collecting Coherence Judgments

To comparatively evaluate the coherence models introduced above, we first needed to establish an independent measure of coherence by eliciting judgments from human participants.

Materials and Design For optimizing and testing our models, we created a corpus representative of coherent and incoherent texts. More specifically, we elicited coherence judgments for multi-document summaries produced by systems and human writers for the Document Understanding Conference (DUC, 2003). Automatically generated summaries are prone to coherence violations [Mani, 2001] — sentences are often extracted out of context and concatenated to form a text — and are therefore a good starting point for coherence analysis.

We randomly selected 16 input document clusters[1] and included in our materials six summaries corresponding to each cluster. One summary was authored by a human, whereas five were produced by automatic summarization systems that participated in DUC. Thus the set of materials contained $6 \cdot 16 = 96$ summaries. From these, six summaries were reserved as a development set, whereas the remaining 90 summaries were used for testing (see Section 5).

Procedure and Subjects Participants first saw a set of instructions that explained the task, defined local coherence, and provided several examples. Then the summaries were presented; a new random order of presentation was generated for each participant. Participants were asked to use a seven point scale to rate how coherent the summaries were without having seen the source texts. The study was conducted remotely over the Internet and was completed by 177 unpaid volunteers (approximately 23 per summary), all native speakers of English.

The judgments were averaged to provide a single rating per summary and all further analyses were conducted on means. We report results on inter-subject agreement in Section 5.2.

5 Experiments

All our models were evaluated on the DUC summaries for which we elicited coherence judgments. We used correlation analysis to assess the degree of linear relationship between human ratings and coherence as estimated by our models. In this section we provide an overview of the parameters we explored, and present and discuss our results.

5.1 Parameter Estimation

Since the models we evaluated have different training requirements (e.g., some must be trained on a large corpus and others are not corpus specific), it was not possible to have a uniform training corpus for all models. Here, we give an overview of the data requirements for our models and explain how these were met in our experiments.

For the entity-based models, we need corpus data to select an appropriate history size and to estimate the probability of various entity transitions. History size was adjusted using the development corpus described above. Entity transition probabilities were estimated from a different corpus: we used 88 human summaries that were produced by DUC analysts. We assumed that human authored summaries were coherent and therefore appropriate for obtaining reliable estimates. The grid columns were augmented with the start and end symbols, increasing the size of the grid cell categories to six. We estimated the probabilities of individual columns using n-gram models (see (3)) of variable length (i.e., 2–4) smoothed with Witten Bell discounting.

[1] Summaries with ungrammatical sentences were excluded from our materials, to avoid eliciting judgments on text properties that were not coherence-related.

	Humans	Egrid	Overlap	LSA	HStO	Lesk	JCon	Lin
Egrid	.246*							
Overlap	.120	−.341**						
LSA	.230*	.042	.013					
HStO	.322**	.071	.093	.037				
Lesk	.125	.227	−.032	.098	.380**			
JCon	−.290**	−.392**	.485**	.035	.625**	.270*		
Lin	.173	.074	−.107	.053	.776**	.421**	.526**	
Resnik	.207	−.003	.052	−.063	.746**	.410**	.606**	.809**
*$p < .05$ (2-tailed) **$p < .01$ (2-tailed)								

Table 4: Correlation between human ratings and coherence models, measured by Pearson coefficient. Stars indicate the level of statistical significance.

Some of the models based on semantic relatedness provide a coherence score without presupposing access to corpus data. These models are based on the measures proposed by Hirst and St-Onge [1998], Lesk [1986] and on word overlap. Since they require no prior training, they were directly computed on a lemmatized version of the test set. Other semantic association models rely on corpus data to either automatically construct representations of lexical meaning [Foltz *et al.*, 1998] or to populate WordNet's representations with frequency counts [Resnik, 1995; Jiang and Conrath, 1997; Lin, 1998]. We used a lemmatized version of the North American News Text Corpus for this purpose. The corpus contains 350 million words of text taken from a variety of news sources and is therefore appropriate for modeling the coherence of news summaries. Vector-based representations were created using a term-document matrix. We used singular value decomposition [Berry *et al.*, 1994] to reduce the vector space to 100 dimensions and thus obtained a semantic space similar to Latent Semantic Analysis (LSA, Foltz *et al.* [1998]).

5.2 Results

The evaluation of our coherence models was driven by two questions: (a) What is the contribution of various linguistic sources to coherence modeling? and (b) Are the two modeling frameworks complementary or redundant?

Upper bound We first assessed human agreement on coherence judgments. Inter-subject agreement gives an upper bound for the task and allows us to interpret model performance in relation to human performance. To measure agreement, we employed leave-one-out resampling [Weiss and Kulikowski, 1991]. The technique correlates the ratings of each participant with the mean ratings of all other participants. The average agreement was $r = .768$.

Model Performance Table 4 displays the results of correlation analysis on 58 summaries[2] from our test set. Significant correlations are observed for the entity grid model (Egrid), the vector-based model (LSA), and two WordNet-based models, employing Hirst and St-Onge's [1998] (HStO) and Jiang and Conrath's [1997] (JCon) similarity measures.

The significant correlation of the Egrid model with human judgments empirically validates Centering's claims about the importance of entity transitions in establishing local coherence. The performance of the LSA model confirms previous claims in the literature [Foltz *et al.*, 1998] that distributional similarity is a significant predictor of text coherence. The fact that JCon correlates[3] with human judgments is not surprising either; its performance has been superior to other WordNet-based measures on two tasks that are particularly relevant for the problem considered here: modeling human similarity judgments and the automatic detection of malapropisms, a phenomenon that often results in locally incoherent discourses (see Budanitsky and Hirst [2001] for details). It is worth considering why HStO performs better than JCon. In quantifying semantic similarity, HStO uses all available relations in WordNet (e.g., antonymy, meronymy, hyponymy) and is therefore better-suited to capture cohesive ties arising from semantically related words than JCon which exploits solely hyponymy. These results further suggest that the chosen similarity measure plays an important role in modeling coherence.

The correlations obtained by our models are substantially lower when compared with the inter-subject agreement of .768. Our results indicate that there is no single method which captures all aspects of local coherence, although HStO yields the highest correlation coefficient in absolute terms. It is worth noting that Egrid is competitive with LSA and the semantic association models based on WordNet, even though it is unlexicalized. We conjecture that Egrid makes up for the lack of lexicalization by taking syntactic information into account and having a more global perspective of discourse. This is achieved by explicitly modeling entity transitions spanning more than two consecutive sentences.

As far as model intercorrelations are concerned, note that Egrid is not significantly correlated with the LSA model, thus indicating that the two models capture complementary coherence properties. Interestingly, LSA displays no correlation with the WordNet-based models. Although both types of models rely on semantic relatedness, they employ distinct representations of lexical meaning (word co-occurrence-based vs. taxonomy-based). Expectedly, the WordNet-based coherence models are all intercorrelated (see Table 4).

Model Combination An obvious question is whether a better fit with the experimental data can be obtained via model combination. A standard way to integrate different predictors (i.e., models) is multiple linear regression. Since several of

[2]The rest of the test data —32 summaries —was used for evaluating our combined model.

[3]The correlation in negative, because JCon is a measure of dissimilarity; the coeffi cient therefore has the opposite sign in comparison with the other WordNet-based measures.

the models were intercorrelated, we performed stepwise regression (using forward selection) to determine the best set of predictors for coherence. The best-fitting combined model (obtained on the set of 58 summaries) included five variables: Egrid, Overlap, LSA, HStO, and Lesk.

Next we tested the combined model's performance on 32 unseen summaries, the output of two systems different from those used for assessing the contribution of the individual models. More specifically, we obtained a coherence score for each unseen summary using the combined model and then compared the estimated values and the human ratings. The comparison yielded a correlation coefficient of .522 ($p < .01$) outperforming any single model. This is a linguistically richer model which integrates several interrelated aspects of coherence (both syntactic and semantic) such as repetition (Overlap), syntactic prominence (Egrid), and semantic association (LSA, HStO, Lesk).

6 Conclusions

In this paper we have compared and contrasted two main frameworks for representing and measuring text coherence. Our syntactic framework operationalizes Centering's notion of local coherence using entity grids. We argued that this representation is particularly suited for uncovering entity transition types typical of coherent and incoherent texts and introduced a model that quantitatively delivers this assessment. Our semantic framework capitalized on the notion of similarity between sentences. We experimented with a variety of similarity measures employing different representations of lexical meaning: word-based, distributional, and taxonomy-based. Our experiments revealed that the two modeling approaches are complementary: our best model retains aspects of entity coherence as well as semantic relatedness.

The modeling approach taken in this paper relies on shallow text properties and is relatively inexpensive (assuming access to a taxonomy and a parser). This makes the proposed models particularly attractive for the automatic evaluation of machine generated summaries. An important future direction lies in the development of a lexicalized version of the entity-grid model that combines the benefits of grid column topology and semantic similarity. Further investigations on different languages, text types, and genres will test the generality and portability of our models. We will also examine whether additional linguistic knowledge (e.g., coreference resolution, causality) will result in improved performance.

Acknowledgments

The authors acknowledge the support of EPSRC (Lapata; grant GR/T04540/01) and the National Science Foundation (Barzilay; CAREER grant IIS-0448168). Thanks to Eli Barzilay, Frank Keller, Smaranda Muresan, Kevin Simler, Caroline Sporleder, Chao Wang, and Bonnie Webber for helpful comments and suggestions.

References

[Bangalore *et al.*, 2000] Srinivas Bangalore, Owen Rambow, and Steven Whittaker. Evaluation metrics for generation. In *Proceedings of the INLG*, pages 1–8, 2000.

[Berry *et al.*, 1994] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595, 1994.

[Budanitsky and Hirst, 2001] Alexander Budanitsky and Graeme Hirst. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proceedings of ACL Worskhop on WordNet and Other Lexical Resources*, pages 29–34, 2001.

[Collins, 1998] Michael Collins. *Head-driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, 1998.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet: An Electronic Database*. MIT Press, 1998.

[Foltz *et al.*, 1998] Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. Textual coherence using latent semantic analysis. *Discourse Processes*, 25(2&3):285–307, 1998.

[Grosz *et al.*, 1995] Barbara Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

[Halliday and Hasan, 1976] M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, 1976.

[Higgins *et al.*, 2004] Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the NAACL*, pages 185-192, 2004.

[Hirst and St-Onge, 1998] Graeme Hirst and David St-Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database and Some of its Applications*, pages 305–332. MIT Press, 1998.

[Hovy and Lin, 2003] Eduard Hovy and Chin-Yew Lin. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the HLT/NAACL*, pages 71–78, 2003.

[Jiang and Conrath, 1997] Jay J. Jiang and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of ROCLING*, 1997.

[Lesk, 1986] Michael Lesk. Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 Special Interest Group in Documentation*, pages 24–26, 1986.

[Lin, 1998] Dekang Lin. An information-theoretic definition of similarity. In *Proceedings of the ICML*, pages 296–304, 1998.

[Mani, 2001] Inderjeet Mani. *Automatic Summarization*. John Benjamins Pub Co, 2001.

[Miltsakaki and Kukich, 2004] Eleni Miltsakaki and Karen Kukich. Evaluation of text coherence for electronic essay scoring systems. natural language engineering. *Natural Language Engineering*, 10(1):25–55, 2004.

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

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLUE: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, pages 311–318, 2002.

[Resnik, 1995] Philip Resnik. Using information content to evaluate semantic similarity. In *Proceedings of 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

[Walker *et al.*, 1998] Marilyn A. Walker, Arvind K. Joshi, and Ellen F. Prince, editors. *Centering Theory in Discourse*. Clarendon Press, Oxford, 1998.

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