Modeling Local Coherence: An Entity-Based Approach

Regina Barzilay*
Massachusetts Institute of Technology

Mirella Lapata**
University of Edinburgh

This article proposes a novel framework for representing and measuring local coherence. Central to this approach is the entity-grid representation of discourse, which captures patterns of entity distribution in a text. The algorithm introduced in the article automatically abstracts a text into a set of entity transition sequences and records distributional, syntactic, and referential information about discourse entities. We re-conceptualize coherence assessment as a learning task and show that our entity-based representation is well-suited for ranking-based generation and text classification tasks. Using the proposed representation, we achieve good performance on text ordering, summary coherence evaluation, and readability assessment.

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 article focuses on **local coherence**, which captures text relatedness at the level of sentence-to-sentence transitions. Local coherence is undoubtedly necessary for **global coherence** and has received considerable attention in computational linguistics (Foltz, Kintsch, and Landauer 1998; Marcu 2000; Lapata 2003; Althaus, Karamanis, and Koller

* Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar Street, 32-G468 Cambridge, MA 02139. E-mail: regina@csail.mit.edu.
** School of Informatics, University of Edinburgh, EH8 9LW, Edinburgh, UK. E-mail: mlap@inf.ed.ac.uk.

Submission received: 29 November 2005; revised submission received: 6 March 2007; accepted for publication: 5 May 2007.

2004; Karamanis et al. 2004). It is also supported by much psycholinguistic evidence. For instance, McKoon and Ratcliff (1992) argue that local coherence is the primary source of inference-making during reading.

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, Joshi, and Weinstein 1995) and other entity-based theories of discourse (e.g., Givon 1987; Prince 1981). The algorithm introduced in the article 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 the proposed entity-based representation of discourse allows us to learn the properties of coherent texts from a corpus, without recourse to manual annotation or a predefined knowledge base. We demonstrate the usefulness of this representation by testing its predictive power in three applications: text ordering, automatic evaluation of summary coherence, and readability assessment.

We formulate the first two problems—text ordering and summary evaluation—as ranking problems, and present an efficiently learnable model that ranks 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, in much the same way speech recognizers employ language models at the sentence level.

In the text-ordering task our algorithm has to select a maximally coherent sentence order from a set of candidate permutations. In the summary evaluation task, we compare the rankings produced by the model against human coherence judgments elicited for automatically generated summaries. In both experiments, our method yields improvements over state-of-the-art models. We also show the benefits of the entity-based representation in a readability assessment task, where the goal is to predict the comprehension difficulty of a given text. In contrast to existing systems which focus on intra-sentential features, we explore the contribution of discourse-level features to this task. By incorporating coherence features stemming from the proposed entity-based representation, we improve the performance of a state-of-the-art readability assessment system (Schwarm and Ostendorf 2005).

In the following section, we provide an overview of entity-based theories of local coherence and outline previous work on its computational treatment. Then, we introduce our entity-based representation, and define its linguistic properties. In the subsequent sections, we present our three evaluation tasks, and report the results of our experiments. Discussion of the results concludes the article.

2. Related Work

Our approach is inspired by entity-based theories of local coherence, and is well-suited for developing a coherence metric in the context of a ranking-based text generation system. We first summarize entity-based theories of discourse, and overview previous attempts for translating their underlying principles into computational coherence models. Next, we describe ranking approaches to natural language generation and focus on coherence metrics used in current text planners.

2.1 Entity-Based Approaches to Local Coherence

Linguistic Modeling. Entity-based accounts of local coherence have a long tradition within the linguistic and cognitive science literature (Kuno 1972; Chafe 1976; Halliday and Hasan 1976; Karttunen 1976; Clark and Haviland 1977; Prince 1981; Grosz, Joshi, and Weinstein 1995). A unifying assumption underlying different approaches is that discourse coherence is achieved in view of the way discourse entities are introduced and discussed. This observation is commonly formalized by devising constraints on the linguistic realization and distribution of discourse entities in coherent texts.

At any point in the discourse, some entities are considered more salient than others, and consequently are expected to exhibit different properties. In Centering Theory (Grosz, Joshi, and Weinstein 1995; Walker, Joshi, and Prince 1998; Strube and Hahn 1999; Poesio et al. 2004), salience concerns how entities are realized in an utterance (e.g., whether they are they pronominalized or not). In other theories, salience is defined in terms of topicality (Chafe 1976; Prince 1978), predictability (Kuno 1972; Halliday and Hasan 1976), and cognitive accessibility (Gundel, Hedberg, and Zacharski 1993). More refined accounts expand the notion of salience from a binary distinction to a scalar one; examples include Prince's (1981) familiarity scale, and Givon's (1987) and Ariel's (1988) givenness-continuum.

The salience status of an entity is often reflected in its grammatical function and the linguistic form of its subsequent mentions. Salient entities are more likely to appear in prominent syntactic positions (such as subject or object), and to be introduced in a main clause. The linguistic realization of subsequent mentions—in particular, pronominalization—is so tightly linked to salience that in some theories (e.g., Givon 1987) it provides the sole basis for defining a salience hierarchy. The hypothesis is that the degree of underspecification in a referring expression indicates the topical status of its antecedent (e.g., pronouns refer to very salient entities, whereas full NPs refer to less salient ones). In Centering Theory, this phenomenon is captured in the *Pronoun Rule*, and Givon's *Scale of Topicality* and Ariel's *Accessibility Marking Scale* propose a graded hierarchy of underspecification that ranges from zero anaphora to full noun phrases, and includes stressed and unstressed pronouns, demonstratives with modifiers, and definite descriptions.

Entity-based theories capture coherence by characterizing the distribution of entities across discourse utterances, distinguishing between salient entities and the rest. The intuition here is that texts about the same discourse entity are perceived to be more coherent than texts fraught with abrupt switches from one topic to the next. The patterned distribution of discourse entities is a natural consequence of topic continuity observed in a coherent text. Centering Theory formalizes fluctuations in topic continuity in terms of transitions between adjacent utterances. The transitions are ranked, that is, texts demonstrating certain types of transitions are deemed more coherent than texts where such transitions are absent or infrequent. For example, CONTINUE transitions require that two utterances have at least one entity in common and are preferred over transitions that repeatedly SHIFT from one entity to the other. Givon's (1987) and Hoey's (1991) accounts of discourse continuity complement local measurements by considering global characteristics of entity distribution, such as the lifetime of an entity in discourse and the referential distance between subsequent mentions.

Computational Modeling. An important practical question is how to translate principles of these linguistic theories into a robust coherence metric. A great deal of research has been devoted to this issue, primarily in Centering Theory (Miltsakaki and Kukich

2000; Hasler 2004; Karamanis et al. 2004). Such translation is challenging in several respects: one has to determine ways of combining the effects of various constraints and to instantiate parameters of the theory that are often left underspecified. Poesio et al. (2004) note that even for fundamental concepts of Centering Theory such as "utterance," "realization," and "ranking," multiple—and often contradictory—interpretations have been developed over the years, because in the original theory these concepts are not explicitly fleshed out. For instance, in some Centering papers, entities are ranked with respect to their grammatical function (Brennan, Friedman, and Pollard 1987; Walker, Iida, and Cote 1994; Grosz, Joshi, and Weinstein 1995), and in others with respect to their position in Prince's (1981) givenness hierarchy (Strube and Hahn 1999) or their thematic role (Sidner 1979). As a result, two "instantiations" of the same theory make different predictions for the same input. Poesio et al. (2004) explore alternative specifications proposed in the literature, and demonstrate that the predictive power of the theory is highly sensitive to its parameter definitions.

A common methodology for translating entity-based theories into computational models is to evaluate alternative specifications on manually annotated corpora. Some studies aim to find an instantiation of parameters that is most consistent with observable data (Strube and Hahn 1999; Karamanis et al. 2004; Poesio et al. 2004). Other studies adopt a specific instantiation with the goal of improving the performance of a metric on a task. For instance, Miltsakaki and Kukich (2000) annotate a corpus of student essays with entity transition information, and show that the distribution of transitions correlates with human grades. Analogously, Hasler (2004) investigates whether Centering Theory can be used in evaluating the readability of automatic summaries by annotating human and machine generated extracts with entity transition information.

The present work differs from these approaches in goal and methodology. Although our work builds upon existing linguistic theories, we do not aim to directly implement or refine any of them in particular. We provide our model with sources of knowledge identified as essential by these theories, and leave it to the inference procedure to determine the parameter values and an optimal way to combine them. From a design viewpoint, we emphasize automatic computation for both the underlying discourse representation and the inference procedure. Thus, our work is complementary to computational models developed on manually annotated data (Miltsakaki and Kukich 2000; Hasler 2004; Poesio et al. 2004). Automatic, albeit noisy, feature extraction allows us to perform a large scale evaluation of differently instantiated coherence models across genres and applications.

2.2 Ranking Approaches in Natural Language Generation

Ranking approaches have enjoyed an increasing popularity at all stages in the generation pipeline, ranging from text planning to surface realization (Knight and Hatzivassiloglou 1995; Langkilde and Knight 1998; Mellish et al. 1998; Walker, Rambow, and Rogati 2001; Karamanis 2003; Kibble and Power 2004). In this framework, an underlying system produces a potentially large set of candidate outputs, with respect to various text generation rules encoded as hard constraints. Not all of the resulting alternatives will correspond to well-formed texts, and of those which may be judged acceptable, some will be preferable to others. The candidate generation phase is followed by an assessment phase in which the candidates are ranked based on a set of desirable properties encoded in a ranking function. The top-ranked candidate is selected for presentation. A two-stage generate-and-rank architecture circumvents the complexity

of traditional generation systems, where numerous, often conflicting constraints, have to be encoded during development in order to produce a single high-quality output.

Because the focus of our work is on text coherence, we discuss here ranking approaches applied to text planning (see Walker et al. [2001] and Knight and Hatzivassiloglou [1995] for ranking approaches to sentence planning and surface realization, respectively). The goal of text planning is to determine the content of a text by selecting a set of information-bearing units and arranging them into a structure that yields well-formed output. Depending on the system, text plans are represented as discourse trees (Mellish et al. 1998) or linear sequences of propositions (Karamanis 2003). Candidate text structures may differ in terms of the selected propositions, the sequence in which facts are presented, the topology of the tree, or the order in which entities are introduced. A set of plausible candidates can be created via stochastic search (Mellish et al. 1998) or by a symbolic text planner following different text-formation rules (Kibble and Power 2004). The best candidate is chosen using an evaluation or ranking function often encoding coherence constraints. Although the type and complexity of constraints vary greatly across systems, they are commonly inspired by Rhetorical Structure Theory or entity-based constraints similar to the ones captured by our method. For instance, the ranking function used by Mellish et al. gives preference to plans where consecutive facts mention the same entities and is sensitive to the syntactic environment in which the entity is first introduced (e.g., in a subject or object position). Karamanis finds that a ranking function based solely on the principle of continuity achieves competitive performance against more sophisticated alternatives when applied to ordering short descriptions of museum artifacts.[1] In other applications, the ranking function is more complex, integrating rules from Centering Theory along with stylistic constraints (Kibble and Power 2004).

A common feature of current implementations is that the specification of the ranking function—feature selection and weighting—is performed manually based on the intuition of the system developer. However, even in a limited domain this task has proven difficult. Mellish et al. (1998; page 100) note: "The problem is far too complex and our knowledge of the issues involved so meager that only a token gesture can be made at this point." Moreover, these ranking functions operate over semantically rich input representations that cannot be created automatically without extensive knowledge engineering. The need for manual coding impairs the portability of existing methods for coherence ranking to new applications, most notably to text-to-text generation applications, such as summarization.

In the next section, we present a method for coherence assessment that overcomes these limitations: We introduce an entity-based representation of discourse that is automatically computed from raw text; we argue that the proposed representation reveals entity transition patterns characteristic of coherent texts. The latter can be easily translated into a large feature space which lends itself naturally to the effective learning of a ranking function, without explicit manual involvement.

3. The Coherence Model

In this section we describe our entity-based representation of discourse. We explain how it is computed and how entity transition patterns are extracted. We also discuss how

1 Each utterance in the discourse refers to at least one entity in the utterance that precedes it.

these patterns can be encoded as feature vectors appropriate for performing coherence-related ranking and classification tasks.

3.1 The Entity-Grid Discourse 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, and the columns correspond to discourse entities. By **discourse entity** we mean a class of coreferent noun phrases (we explain in Section 3.3 how coreferent entities are identified). For each occurrence of a discourse entity in the text, the corresponding grid cell contains information about its presence or absence in a sequence of sentences. In addition, for entities present in a given sentence, grid cells contain information about their syntactic role. Such information can be expressed in many ways (e.g., using constituent labels or thematic role information). Because grammatical relations figure prominently in entity-based theories of local coherence (see Section 2), they serve as a logical point of departure. Each grid cell thus corresponds to a string from a set of categories reflecting whether the entity in question is a subject (S), object (O), or neither (X). Entities absent from a sentence are signaled by gaps (–). Grammatical role information can be extracted from the output of a broad-coverage dependency parser (Lin 2001; Briscoe and Carroll 2002) or any state-of-the art statistical parser (Collins 1997; Charniak 2000). We discuss how this information was computed for our experiments in Section 3.3.

Table 1 illustrates a fragment of an entity grid constructed for the text in Table 2. Because 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. Also note that the grid in Table 1 takes coreference resolution into account. Even though the same entity appears in different linguistic forms, for example, *Microsoft Corp.*, *Microsoft*, and *the company*, it is mapped to a single entry in the grid (see the column introduced by *Microsoft* in Table 1).

Table 1
A fragment of the entity grid. Noun phrases are represented by their head nouns. Grid cells correspond to grammatical roles: subjects (S), objects (O), or neither (X).

	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 2
Summary augmented with syntactic annotations for grid computation.

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

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).

3.2 Entity Grids as Feature Vectors

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 the 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* and *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 $\{\mathbf{S}, \mathbf{O}, \mathbf{X}, \mathbf{-}\}^n$ 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 can now go one step further and represent each text by a fixed set of transition sequences using a standard feature vector notation. Each grid rendering j of a document d_i corresponds to 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}. This feature vector representation is usefully amenable to machine learning algorithms (see our experiments in Sections 4–6). Furthermore, it allows the consideration of large numbers of transitions which could potentially uncover novel entity distribution patterns relevant for coherence assessment or other coherence-related tasks.

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 second row (introduced by d_1) is the feature vector representation of the grid in Table 1.

3.3 Grid Construction: Linguistic Dimensions

One of the central research issues in developing entity-based models of coherence is determining what sources of linguistic knowledge are essential for accurate prediction, and how to encode them succinctly in a discourse representation. Previous approaches tend to agree on the features of entity distribution related to local coherence—the disagreement lies in the way these features are modeled.

Our study of alternative encodings is not a mere duplication of previous efforts (Poesio et al. 2004) that focus on linguistic aspects of parameterization. Because we are interested in an automatically constructed model, we have to take into account computational and learning issues when considering alternative representations. Therefore, our exploration of the parameter space is guided by three considerations: the linguistic importance of a parameter, the accuracy of its automatic computation, and the size of the resulting feature space. From the linguistic side, we focus on properties of entity distribution that are tightly linked to local coherence, and at the same time allow for multiple interpretations during the encoding process. Computational considerations prevent us from considering discourse representations that cannot be computed reliably by existing tools. For instance, we could not experiment with the granularity of an utterance—sentence versus clause—because available clause separators introduce substantial noise into a grid construction. Finally, we exclude representations that will explode the size of the feature space, thereby increasing the amount of data required for training the model.

Entity Extraction. The accurate computation of entity classes is key to computing meaningful entity grids. In previous implementations of entity-based models, classes of coreferent nouns have been extracted manually (Miltsakaki and Kukich 2000; Karamanis et al. 2004; Poesio et al. 2004), but this is not an option for our model. An obvious solution for identifying entity classes is to employ an automatic coreference resolution tool that determines which noun phrases refer to the same entity in a document.

Current approaches recast coreference resolution as a classification task. A pair of NPs is classified as coreferring or not based on constraints that are learned from an annotated corpus. A separate clustering mechanism then coordinates the possibly contradictory pairwise classifications and constructs a partition on the set of NPs. In our experiments, we employ Ng and Cardie's (2002) coreference resolution system. The system decides whether two NPs are coreferent by exploiting a wealth of lexical, grammatical, semantic, and positional features. It is trained on the MUC (6–7) data sets and yields state-of-the-art performance (70.4 F-measure on MUC-6 and 63.4 on MUC-7).

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

	SS	SO	SX	S–	OS	OO	OX	O–	XS	XO	XX	X–	–S	–O	–X	––
d_1	.01	.01	0	.08	.01	0	0	.09	0	0	0	.03	.05	.07	.03	.59
d_2	.02	.01	.01	.02	0	.07	0	.02	.14	.14	.06	.04	.03	.07	0.1	.36
d_3	.02	0	0	.03	.09	0	.09	.06	0	0	0	.05	.03	.07	.17	.39

Although machine learning approaches to coreference resolution have been reasonably successful—state-of-the-art coreference tools today reach an F-measure[2] of 70% when trained on newspaper texts—it is unrealistic to assume that such tools will be readily available for different domains and languages. We therefore consider an additional approach to entity extraction where entity classes are constructed 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. Under this view *Microsoft Corp.* from Table 2 (Sentence 1) corresponds to two entities, *Microsoft* and *Corp.*, which are in turn distinct from *the company*. This approach is only a rough approximation to fully fledged coreference resolution, but it is simple from an implementational perspective and produces consistent results across domains and languages.

Grammatical Function. Several entity-based approaches assert that grammatical function is indicative of an entity's prominence in discourse (Hudson, Tanenhaus, and Dell 1986; Kameyama 1986; Brennan, Friedman, and Pollard 1987; Grosz, Joshi, and Weinstein 1995). Most theories discriminate between subject, object, and the remaining grammatical roles: subjects are ranked higher than objects, and these are ranked higher than other grammatical functions.

In our framework, we can easily assess the impact of syntactic knowledge by modifying how transitions are represented in the entity grid. In syntactically aware grids, transitions are expressed by four categories: **S**, **O**, **X** and **–**, whereas in simplified grids, we only record whether an entity is present (**X**) or absent (**–**) in a sentence.

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. The phrase-structure output of Collins's parser is transformed into a dependency tree from which grammatical relations are extracted. 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). For more details on the grammatical relations extraction component we refer the interested reader to Barzilay (2003).

Salience. Centering and other discourse theories conjecture that the way an entity is introduced and mentioned depends on its global role in a given discourse. We evaluate the impact of salience information by considering two types of models: The first model treats all entities uniformly, whereas the second one discriminates between transitions of salient entities and the rest. We identify salient entities based on their frequency,[3] following the widely accepted view that frequency of occurrence correlates with discourse prominence (Givon 1987; Ariel 1988; Hoey 1991; Morris and Hirst 1991).

To implement a salience-based model, we modify our feature generation procedure by computing transition probabilities for each salience group separately, and then

2 When evaluating the output of coreference algorithms, performance is typically measured using a model-theoretic scoring scheme proposed in Vilain et al. (1995). The scoring algorithm computes the recall error by taking each equivalence class S in the gold standard and determining the number of coreference links m that would have to be added to the system's output to place all entities in S into the same equivalence class produced by the system. Recall error then is the sum of ms divided by the number of links in the gold standard. Precision error is computed by reversing the roles of the gold standard and system output.

3 The frequency threshold is empirically determined on the development set. See Section 4.2 for further discussion.

combining them into a single feature vector. For n transitions with k salience classes, the feature space will be of size $n \times k$. While we can easily build a model with multiple salience classes, we opt for a binary distinction (i.e., $k = 2$). This is more in line with theoretical accounts of salience (Chafe 1976; Grosz, Joshi, and Weinstein 1995) and results in a moderate feature space for which reliable parameter estimation is possible. Considering a large number of salience classes would unavoidably increase the number of features. Parameter estimation in such a space requires a large sample of training examples that is unavailable for most domains and applications.

Different classes of models can be defined along the linguistic dimensions just discussed. Our experiments will consider several models with varying degrees of linguistic complexity, while attempting to strike a balance between expressivity of representation and ease of computation. In the following sections we evaluate their performance on three tasks: sentence ordering, summary coherence rating, and readability assessment.

3.4 Learning

Equipped with the feature vector representation introduced herein, we can view coherence assessment as a machine learning problem. When considering text generation applications, it is desirable to rank rather than classify instances: There is often no single coherent rendering of a given text but many different possibilities that can be partially ordered. It is therefore not surprising that systems often employ scoring functions to select the most coherent output among alternative renderings (see the discussion in Section 2.2). In this article we argue that encoding texts as entity transition sequences constitutes an appropriate feature set for *learning* (rather than manually specifying) such a ranking function (see Section 4 for details). We present two task-based experiments that put this hypothesis to the test: information ordering (Experiment 1) and summary coherence rating (Experiment 2). Both tasks can be naturally formulated as ranking problems; the learner 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 are a set of different sentence orderings of the same text and a set of summaries produced by different systems for the same document. Note that in both ranking experiments we assume that the algorithm is provided with a limited number of alternatives. In practice, the space of candidates can be vast, and finding the optimal candidate may require pairing our ranking algorithm with a decoder similar to the ones used in machine translation (Germann et al. 2004).

Although the majority of our experiments fall within the generate-and-rank framework previously sketched, nothing prevents the use of our feature vector representation for conventional classification tasks. We offer an illustration in Experiment 3, where features extracted from entity grids are used to enhance the performance of a readability assessment system. Here, the learner takes as input a set of documents labeled with discrete classes (e.g., denoting whether a text is difficult or easy to read) and learns to make predictions for unseen instances (see Section 6 for details on the machine learning paradigm we employ).

4. Experiment 1: Sentence 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. The information bearing items can be database entries (Karamanis et al. 2004), propositions (Mellish et al. 1998) or sentences (Lapata 2003; Barzilay and Lee 2004). In sentence ordering, a document is viewed as a bag of sentences and the algorithm's task is to try to find the ordering which maximizes coherence according to some criterion (e.g., the probability of an order).

As explained previously, we use our coherence model to rank alternative sentence orderings instead of trying to find an optimal ordering. We do not assume that local coherence is sufficient to uniquely determine a maximally coherent ordering—other constraints clearly play a role here. It is nevertheless a key property of well-formed text (documents lacking local coherence are naturally globally incoherent), and a model which takes it into account should be able to discriminate coherent from incoherent texts. In our sentence-ordering task we generate random permutations of a test document and measure how often a permutation is ranked higher than the original document. A non-deficient model should prefer the original text more frequently than its permutations (see Section 4.2 for details).

We begin by explaining how a ranking function can be learned for the sentence ordering task. Next, we give details regarding the corpus used for our experiments, describe the methods used for comparison with our approach, and note the evaluation metric employed for assessing model performance. Our results are presented in Section 4.3.

4.1 Modeling

Our training set consists of ordered pairs of alternative renderings (x_{ij}, x_{ik}) of the same document d_i, where x_{ij} exhibits a higher degree of coherence than x_{ik} (we describe in Section 4.2 how such training instances are obtained). Without loss of generality, we assume $j > k$. The goal of the training procedure is to find a parameter vector $\mathbf{w}$ that yields a "ranking score" function which minimizes the number of violations of pairwise rankings provided in the training set

$$\forall (x_{ij}, x_{ik}) \in r^* : \mathbf{w} \cdot \Phi(x_{ij}) > \mathbf{w} \cdot \Phi(x_{ik})$$

where $(x_{ij}, x_{ik}) \in r^*$ if x_{ij} is ranked higher than x_{ik} for the optimal ranking r^* (in the training data), and $\Phi(x_{ij})$ and $\Phi(x_{ik})$ are a mapping onto features representing the coherence properties of renderings x_{ij} and x_{ik}. In our case the features correspond to the entity transition probabilities introduced in Section 3.2. Thus, the ideal ranking function, represented by the weight vector $\mathbf{w}$ would satisfy the condition

$$\mathbf{w} \cdot (\Phi(x_{ij}) - \Phi(x_{ik})) > 0 \; \forall j, i, k \text{ such that } j > k$$

The problem is typically treated as a Support Vector Machine constraint optimization problem, and can be solved using the search technique described in Joachims (2002). This approach has been shown to be highly effective in various tasks ranging from collaborative filtering (Joachims 2002) to parsing (Toutanova, Markova, and Manning 2004). Other discriminative formulations of the ranking problem are possible (Collins 2002; Freund et al. 2003); however, we leave this to future work.

Table 4
The size of the training and test instances for the Earthquakes and Accidents corpora (measured by the number of pairs that contain the original order and a random permutation of this order).

	Training	Testing
Earthquakes	1,896	2,056
Accidents	2,095	2,087

Once the ranking function is learned, unseen renderings (x_{ij}, x_{ik}) of document d_i can be ranked simply by computing the values $\mathbf{w}^*\Phi(x_{ij})$ and $\mathbf{w}^*\Phi(x_{ik})$ and sorting them accordingly. Here, $\mathbf{w}^*$ is the optimized parameter vector resulting from training.

4.2 Method

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 enables large-scale automatic evaluation and is widely used in assessing ordering algorithms (Karamanis 2003; Lapata 2003; Althaus, Karamanis, and Koller 2004; Barzilay and Lee 2004). 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. Because 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 herein, we collected data[4] 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 earthquakes (Earthquakes). The second includes narratives from the National Transportation Safety Board's aviation accident database (Accidents). Both corpora 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 up to 20 randomly generated permutations for training.[5] A similar method was used to obtain the test data. Table 4 shows the size of the training and test corpora used in our experiments. We held out 10 documents (i.e., 200 pairwise rankings) from the training data for development purposes.

Features and Parameter Settings. In order to investigate the contribution of linguistic knowledge on model performance we experimented with a variety of grid representations resulting in different parameterizations of the feature space from which our model is learned. We focused on three sources of linguistic knowledge—syntax, coreference resolution, and salience—which play a prominent role in entity-based analyses of dis-

4 The collections are available from `http://people.csail.mit.edu/regina/coherence/`.
5 Short texts may have less than 20 permutations. The corpus described in the original ACL publication (Barzilay and Lapata 2005) contained a number of duplicate permutations. These were removed from the current version of the corpus.

course coherence (see Section 3.3 for details). An additional motivation for our study was to explore 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.

We thus compared a linguistically rich model against models that use more impoverished representations. More concretely, our full model (**Coreference**+**Syntax**+ **Salience**+) uses coreference resolution, denotes entity transition sequences via grammatical roles, and differentiates between salient and non-salient entities. Our less-expressive models (seven in total) use only a subset of these linguistic features during the grid construction process. We evaluated the effect of syntactic knowledge by eliminating the identification of grammatical relations and recording solely whether an entity is present or absent in a sentence. This process created a class of four models of the form **Coreference[**+/−**]Syntax**−**Salience[**+/−**]**. The effect of fully fledged coreference resolution was assessed by creating models where entity classes were constructed simply by clustering nouns on the basis of their identity (**Coreference**−**Syntax[**+/−**]Salience[**+/−**]**). Finally, the contribution of salience was measured by comparing the full model which accounts separately for patterns of salient and non-salient entities against models that do not attempt to discriminate between them (**Coreference[**+/−**]Syntax[**+/−**]Salience**−).

We would like to note that in this experiment we apply a coreference resolution tool to the original text and then generate permutations for the pairwise ranking task. An alternative design is to apply coreference resolution to permuted texts. Because existing methods for coreference resolution take into consideration the order of noun phrases in a text, the accuracy of these tools on permuted sentence sequences is close to random. Therefore, we opt to resolve coreference within the original text. Although this design has an oracle feel to it, it is not uncommon in practical applications. For instance, in text generation systems, content planners often operate over fully specified semantic representations, and can thus take advantage of coreference information during sentence ordering.

Besides variations in the underlying linguistic representation, our model is also specified by 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. Optimal salience-based models were obtained for entities with frequency $\geq$2. The optimal transition length was $\leq$3.[6]

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

Comparison with State-of-the-Art Methods. We compared the performance of our algorithm against two state-of-the-art models proposed by Foltz, Kintsch, and Landauer (1998) and Barzilay and Lee (2004). These models rely largely on lexical information for assessing document coherence, contrary to our models which are in essence unlexicalized. Recall from Section 3 that our approach captures local coherence by modeling patterns of entity distribution in discourse, without taking note of their lexical instantiations. In the following we briefly describe the lexicalized models we employed in our comparative study and motivate their selection.

6 The models we used in our experiments are available from `http://people.csail.mit.edu/regina/coherence/` and `http://homepages.inf.ed.ac.uk/mlap/coherence/`.

Foltz, Kintsch, and Landauer (1998) model measures coherence as a function of semantic relatedness between adjacent sentences. The underlying intuition here is that coherent texts will contain a high number of semantically related words. Semantic relatedness is computed automatically using Latent Semantic Analysis (LSA; Landauer and Dumais 1997) from raw text without employing syntactic or other annotations. In this framework, a word's meaning is captured in a multi-dimensional space by a vector representing its co-occurrence with neighboring words. Co-occurrence information is collected in a frequency matrix, where each row corresponds to a unique word, and each column represents a given linguistic context (e.g., sentence, document, or paragraph). Foltz, Kintsch, and Landauer's model use singular value decomposition (SVD; Berry, Dumais, and O'Brien 1994) to reduce the dimensionality of the space. The transformation renders sparse matrices more informative and can be thought of as a means of uncovering latent structure in distributional data. The meaning of a sentence is next represented as a vector by taking the mean of the vectors of its words. The similarity between two sentences is determined by measuring the cosine of their means:

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

where $\mu(\vec{S_i}) = \frac{1}{|S_i|}\sum_{\vec{u}\in S_i}\vec{u}$, and $\vec{u}$ is the vector for word u. An overall text coherence measure can be easily obtained by averaging the cosines for all pairs of adjacent sentences S_i and S_{i+1}:

$$\text{coherence}(T) = \frac{\sum_{i=1}^{n-1}\cos(S_i, S_{i+1})}{n-1} \tag{2}$$

This model is a good point of comparison for several reasons: (a) it is fully automatic and has relatively few parameters (i.e., the dimensionality of the space and the choice of similarity function), (b) it correlates reliably with human judgments and has been used to analyze discourse structure, and (c) it models an aspect of local coherence which is orthogonal to ours. The LSA model is lexicalized: coherence amounts to quantifying the degree of semantic similarity between sentences. In contrast, our model does not incorporate any notion of similarity: coherence is encoded in terms of transition sequences that are document-specific rather than sentence-specific.

Our implementation of the LSA model followed closely Foltz, Kintsch, and Landauer (1998). We constructed vector-based representations for individual words from a lemmatized version of the North American News Corpus[7] (350 million words) using a term–document matrix. We used SVD to reduce the semantic space to 100 dimensions obtaining thus a space similar to LSA. We estimated the coherence of a document using Equations (1) and (2). A ranking can be trivially inferred by comparing the

7 Our selection of this corpus was motivated by two factors: (a) the corpus is large enough to yield a reliable semantic space, and (b) it consists of news stories and is therefore similar in style, vocabulary, and content to most of the corpora employed in our coherence experiments.

coherence score assigned to the original document against each of its permutations. Ties are resolved randomly.

Both LSA and our entity-grid model are local—they model sentence-to-sentence transitions without being aware of global document structure. In contrast, the content models developed by Barzilay and Lee (2004) learn to represent more global text properties by capturing topics and the order in which these topics appear in texts from the same domain. For instance, a typical earthquake newspaper report contains information about the quake's epicenter, how much it measured, the time it was felt, and whether there were any victims or damage. By encoding constraints on the ordering of these topics, content models have a pronounced advantage in modeling document structure because they can learn to represent how documents begin and end, but also how the discourse shifts from one topic to the next. Like LSA, the content models are lexicalized; however, unlike LSA, they are domain-specific, and would expectedly yield inferior performance on out-of-domain texts.

Barzilay and Lee (2004) implemented content models using an HMM wherein states correspond to distinct topics (for instance, the epicenter of an earthquake or the number of victims), and state transitions represent the probability of changing from one topic to another, thereby capturing possible topic-presentation orderings within a domain. Topics refer to text spans of varying granularity and length. Barzilay and Lee used sentences in their experiments, but clauses or paragraphs would also be possible.

Barzilay and Lee (2004) employed their content models to find a high-probability ordering for a document whose sentences had been randomly shuffled. Here, we use content models for the simpler coherence ranking task. Given two text permutations, we estimate their likelihood according to their HMM model and select the text with the highest probability. Because the two candidates contain the same set of sentences, the assumption is that a more probable text corresponds to an ordering that is more typical for the domain of interest.

In our experiments, we built two content models, one for the Accidents corpus and one for the Earthquake corpus. Although these models are trained in an unsupervised fashion, a number of parameters related to the model topology (i.e., number of states and smoothing parameters) affect their performance. These parameters were tuned on the development set and chosen so as to optimize the models' performance on the pairwise ranking task.

Evaluation Metric. Given a set of pairwise rankings (an original document and one of its permutations), 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%.

4.3 Results

Impact of Linguistic Representation. We first investigate how different types of linguistic knowledge influence our model's performance. Table 5 shows the accuracy on the ordering task when the model is trained on different grid representations. As can be seen, in both domains, the full model **Coreference**+**Syntax**+**Salience**+ significantly outperforms a linguistically naive model which simply records the presence (and absence) of entities in discourse (**Coreference**−**Syntax**−**Salience**−). Moreover, we observe that linguistically impoverished models consistently perform worse than their linguistically elaborate counterparts. We assess whether differences in accuracy are statistically

Table 5
Accuracy measured as a fraction of correct pairwise rankings in the test set. **Coreference[+/−]** indicates whether coreference information has been used in the construction of the entity grid. Similarly, **Syntax[+/−]** and **Salience[+/−]** reflect the use of syntactic and salience information. Diacritics ** ($p < .01$) and * ($p < .05$) indicate whether differences in accuracy between the full model (**Coreference**+**Syntax**+**Salience**+) and all other models are significant (using a Fisher Sign test).

Model	Earthquakes	Accidents
Coreference+**Syntax**+**Salience**+	**87.2**	**90.4**
Coreference+Syntax+Salience−	88.3	90.1
Coreference+Syntax−Salience+	86.6	88.4**
Coreference−Syntax+Salience+	83.0**	89.9
Coreference+Syntax−Salience−	86.1	89.2
Coreference−Syntax+Salience−	82.3**	88.6*
Coreference−Syntax−Salience+	83.0**	86.5**
Coreference−Syntax−Salience−	81.4**	86.0**
HMM-based Content Models	88.0	75.8**
Latent Semantic Analysis	81.0**	87.3**

significant using a Fisher Sign Test. Specifically, we compare the full model against each of the less expressive models (see Table 5).

Let us first discuss in more detail how the contribution of different knowledge sources varies across domains. On the Earthquakes corpus every model that does not use coreference information (**Coreference**−**Syntax[+/−]Salience[+/−]**) performs significantly worse than models augmented with coreference (**Coreference**+ **Syntax[+/−]Salience[+/−]**). This effect is less pronounced on the Accidents corpus, especially for model **Coreference**−**Syntax**+**Salience**+ whose accuracy drops only by 0.5% (the difference between Coreference−Syntax+Salience+ and Coreference+ Syntax+Salience+ is not statistically significant). The same model's performance decreases by 4.2% on the Earthquakes corpus. This variation can be explained by differences in entity realization between the two domains. In particular, the two corpora vary in the amount of coreference they employ; texts from the Earthquakes corpus contain many examples of referring expressions that our simple identity-based approach cannot possibly resolve. Consider for instance the text in Table 6. Here, the expressions *the same area*, *the remote region*, and *site* all refer to *Menglian county*. In comparison, the text from the Accidents corpus contains fewer referring expressions, in fact entities are often repeated verbatim across several sentences, and therefore could be straightforwardly resolved with a shallow approach (see *the pilot*, *the pilot*, *the pilot* in Table 6).

The omission of syntactic information causes a drop in accuracy for models applied to the Accidents corpus. This effect is less noticeable on the Earthquakes corpus (compare the performance of model **Coreference**+**Syntax**−**Salience**+ on the two corpora). We explain this variation by the substantial difference in the type/token ratio between the two domains—12.1 for Earthquakes versus 5.0 for Accidents. The low type/token ratio for Accidents means that most sentences in a text have some words in common. For example, the entities *pilot*, *airplane*, and *airport* appear in multiple sentences in the text from Table 6. Because there is so much repetition in this domain, the syntax-free grids will be relatively similar for both coherent (original) and incoherent texts (permutations). In fact, inspection of the grids from the Accidents corpus reveals that they have many sequences of the form [x x x], [x − − x], [x x − −], and [− − x x] in common,

Table 6
Two texts from the Earthquakes and Accidents corpus. One entity class for each document is shown to demonstrate the difference in referring expressions used in the two corpora.

Example Text from Earthquakes
A strong earthquake hit the China-Burma border early Wednesday morning, but there were no reports of deaths, according to China's Central Seismology Bureau. The 7.3 quake hit [Menglian county] at 5:46 am. [The same area] was struck by a 6.2 temblor early Monday morning, the bureau said. The county is on the China-Burma border, and is a sparsely populated, mountainous region. The bureau's Xu Wei said some buildings sustained damage and there were some injuries, but he had no further details. Communication with [the remote region] is difficult, and satellite phones sent from the neighboring province of Sichuan have not yet reached [the site.] However, he said the likelihood of deaths was low because residents should have been evacuated from [the area] following Monday's quake.
Example Text from Accidents
When [the pilot] failed to arrive for his brother's college graduation, concerned family members reported that he and his airplane were missing. A search was initiated, and the Civil Air Patrol located the airplane on top of Pine Mountain. According to [the pilot]'s flight log, the intended destination was Pensacola, FL, with intermediate stops for fuel at Thomson, GA, and Greenville, AL. Airport personal at Thomson confirmed that the airplane landed about 1630 on 11/6/97. They reported that [the pilot] purchased 26.5 gallons of 100LL fuel and departed about 1700. Witnesses at the Thomson Airport stated that when he took off, the weather was marginal VFR and deteriorating rapidly. Witnesses near Pine Mountain stated that the visibility at the time of the accident was about 1/4 mile in haze/fog.

whereas such sequences are more common in coherent Earthquakes documents and more sparse in their permutations. This indicates that syntax-free analysis can sufficiently discriminate coherent from incoherent texts in the Earthquakes domain, while a more refined representation of entity transition types is required for the Accidents domain.

The contribution of salience is less pronounced in both domains—the difference in performance between the full model (**Coreference**+**Syntax**+**Salience**+) and its salience-agnostic counterpart (**Coreference**+**Syntax**+**Salience**+) is not statistically significant. Salience-based models do deliver some benefits for linguistically impoverished models—for instance, **Coreference**−**Syntax**−**Salience**+ improves over **Coreference**−**Syntax**−**Salience**− ($p < 0.06$) on the Earthquakes corpus. We hypothesize that the small contribution of salience is related to the way it is currently represented. Addition of this knowledge source to our grid representation, doubles the number of features that serve as input to the learning algorithm. In other words, salience-aware models need to learn twice as many parameters as salience-free models, while having access to the same amount of training data. Achieving any improvement in these conditions is challenging.

Comparison with State-of-the-Art Methods. We next discuss the performance of the HMM-based content models (Barzilay and Lee 2004) and LSA (Foltz, Kintsch, and Landauer 1998) in comparison to our model (**Coreference**+**Syntax**+**Salience**+).

First, note that the entity-grid model significantly outperforms LSA on both domains ($p < .01$ using a Sign test, see Table 5). In contrast to our model, LSA is neither entity-based nor unlexicalized: It measures the degree of semantic overlap across successive sentences, without handling discourse entities in a special way (all content words in a sentence contribute towards its meaning). We attribute our model's superior performance, despite the lack of lexicalization, to three factors: (a) the use of more elaborate linguistic knowledge (coreference and grammatical role information); (b) a more holistic representation of coherence (recall that our entity grids operate over texts rather than individual sentences; furthermore, entity transitions can span more than two consecutive sentences, something which is not possible with the LSA model); and (c) exposure to domain relevant texts (the LSA model used in our experiments was not particularly tuned to the Earthquakes or Accidents corpus). Our semantic space was created from a large news corpus (see Section 4.2) covering a wide variety of topics and writing styles. This is necessary for constructing robust vector representations that are not extremely sparse. We thus expect the grid models to be more sensitive to the discourse conventions of the training/test data.

The accuracy of the HMM-based content modes is comparable to the grid model on the Earthquakes corpus (the difference is not statistically significant) but is significantly lower on the Accidents texts (see Table 5). Although the grid model yields similar performance on the two domains, content models exhibit high variability. These results are not surprising. The analysis presented in Barzilay and Lee (2004) shows that the Earthquakes texts are quite formulaic in their structure, following the editorial style of the Associated Press. In contrast, the Accidents texts are more challenging for content models—reports in this set do not undergo centralized editing and therefore exhibit more variability in lexical choice and style. The LSA model also significantly outperforms the content model on the Earthquakes domain ($p < .01$ using a Sign test). Being a local model, LSA is less sensitive to the way documents are structured and is therefore more likely to deliver consistent performance across domains.

The comparison in Table 5 covers a broad spectrum of coherence models. At one end of the spectrum is LSA, a lexicalized model of local discourse coherence which is fairly robust and domain independent. In the middle of the spectrum lies our entity-grid model, which is unlexicalized but linguistically informed and goes beyond simple sentence-to-sentence transitions without, however, fully modeling global discourse structure. At the other end of the spectrum are the HMM-based content models, which are both global and lexicalized. Our results indicate that these models are complementary and that their combination could yield improved results. For example, we could lexicalize our entity grids or supply the content models with local information either in the style of LSA or as entity transitions. However, we leave this to future work.

Training Requirements. We now examine in more detail the training requirements for the entity-grid models. Although for our ordering experiments we obtained training data cheaply, this will not generally be the case and some effort will have to be invested in collecting appropriate data with coherence ratings. We thus address two questions: (1) How much training data is required for achieving satisfactory performance? (2) How domain sensitive are the entity-grid models? In other words, does their performance degrade gracefully when applied to out-of-domain texts?

Figure 1 shows learning curves for the best performing model (**Coreference**+ **Syntax**+**Salience**+) on the Earthquakes and Accidents corpora. We observe that the amount of data required depends on the domain at hand. The Accidents texts are more repetitive and therefore less training data is required to achieve good performance. The

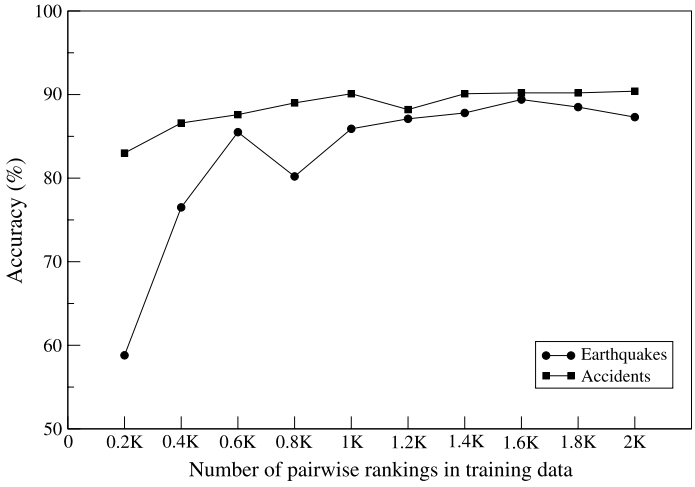

Figure 1
Learning curves for the entity-based model **Coreference**+**Syntax**+**Salience**+ on the Earthquakes and Accidents corpora.

learning curve is steeper for the Earthquakes documents. Irrespective of the domain differences, the model reaches good accuracies when half of the data set is used (1,000 pairwise rankings). This is encouraging, because for some applications (e.g., summarization) large amounts of training data may be not readily available.

Table 7 illustrates the accuracy of the best performing model **Coreference**+ **Syntax**+**Salience**+ when trained on the Earthquakes corpus and tested on Accidents texts and reversely when trained on the Accident corpus and tested on Earthquakes documents. We also illustrate how this model performs when trained and tested on a data set that contains texts from both domains. For the latter experiment the training data set was created by randomly sampling 50 Earthquakes and 50 Accidents documents.

Table 7
Accuracy of entity-based model (Coreference+Syntax+Salience+) and HHM-based content model on out-of-domain texts. Diacritics ** ($p < .01$) and * ($p < .05$) indicate whether performances on in-domain and out-of-domain data are significantly different using a Fisher Sign Test.

Coreference+Syntax+Salience		
Train \ Test	Earthquakes	Accidents
Earthquakes	**87.3**	67.0**
Accidents	69.7**	**90.4**
EarthAccid	86.7	88.5*

HMM-Based Content Models		
Train \ Test	Earthquakes	Accidents
Earthquakes	**88.0**	31.7**
Accidents	60.3**	**75.8**

As can be seen from Table 7, the model's performance degrades considerably (approximately by 20%) when tested on out-of-domain texts. On the positive side, the model's out-of-domain performance is better than chance (i.e., 50%). Furthermore, once the model is trained on data representative of both domains, it performs almost as well as a model which has been trained exclusively on in-domain texts (see the row EarthAccid in Table 7). To put these results into context, we also considered the cross-domain performance of the content models. As Table 7 shows, the decrease in performance is more dramatic for the content models. In fact, the model trained on the Earthquakes domain plummets below the random baseline when applied to the Accidents domain. These results are expected for content models—the two domains have little overlap in topics and do not share structural constraints. Note that the LSA model is not sensitive to cross-domain issues. The semantic space is constructed over many different domains without taking into account style or writing conventions.

The cross-training performance of the entity-based models is somewhat puzzling: these models are not lexicalized, and one would expect that valid entity transitions are preserved across domains. Although transition types are not domain-specific, their distribution could vary from one domain to another. To give a simple example, some domains will have more entities than others (e.g., descriptive texts). In other words, entity transitions capture not only text coherence properties, but also reflect stylistic and genre-specific discourse properties. This hypothesis is indirectly confirmed by the observed differences in the contribution of various linguistic features across the two domains discussed above. Cross-domain differences in the distribution and occurrence of entities have been also observed in other empirical studies of local coherence. For instance, Poesio et al. (2004) show differences in transition types between instructional texts and descriptions of museum texts. In Section 6, we show that features derived from the entity grid help determine the readability level for a given text, thereby verifying more directly the hypothesis that the grid representation captures stylistic discourse factors.

The results presented so far suggest that adapting the proposed model to a new domain would involve some effort in collecting representative texts with associated coherence ratings. Thankfully, the entity grids are constructed in a fully automatic fashion, without requiring manual annotation. This contrasts with traditional implementations of Centering Theory that operate over linguistically richer representations that are typically hand-coded.

5. Experiment 2: Summary Coherence Rating

We further test the ability of our method to assess coherence by comparing model induced rankings against rankings elicited by human judges. Admittedly, the synthetic data used in the ordering task only partially approximates coherence violations that human readers encounter in machine generated texts. 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, because they focus on content similarity between system output and reference texts.

5.1 Modeling

Summary coherence rating can be also formulated as a ranking learning task. We are assuming that the learner has access to several summaries corresponding to the same document or document cluster. Such summaries can be produced by several systems that operate over identical inputs or by a single system (e.g., by varying the compression length or by switching on or off individual system modules, for example a sentence compression or anaphora resolution module). Similarly to the sentence ordering task, our training data includes pairs of summaries (x_{ij}, x_{ik}) of the same document(s) d_i, where x_{ij} is more coherent than x_{ik}. An optimal learner should return a ranking r^* that orders the summaries according to their coherence. As in Experiment 1 we adopt an optimization approach and follow the training regime put forward by Joachims (2002).

5.2 Method

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 has been rated in terms of coherence. One stumbling block to performing this kind of evaluation is the coherence ratings themselves, which are not routinely provided by DUC summary evaluators. In DUC 2003, the quality of automatically generated summaries was assessed along several dimensions ranging from grammatically, to content selection, fluency, and readability. Coherence was indirectly evaluated by noting the number of sentences indicating an awkward time sequence, suggesting a wrong cause–effect relationship, or being semantically incongruent with their neighboring sentences.[8] Unfortunately, the observed coherence violations were not fine-grained enough to be of use in our rating experiments. In the majority of cases DUC evaluators noted either 0 or 1 violations; however, without judging the coherence of the summary as a whole, we cannot know whether a single violation disrupts coherence severely or not.

We therefore obtained judgments for automatically generated summaries from human subjects.[9] We randomly selected 16 input document clusters and five systems that had produced summaries for these sets, along with reference summaries composed by humans. Coherence ratings were collected 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

8 See question 12 in `http://duc.nist.gov/duc2003/quality.html`.
9 The ratings are available from `http://homepages.inf.ed.ac.uk/mlap/coherence/`.

resampling[10] (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 agreement was $r = .768$ ($p < .01$.) Second, we examined the effect of different types of summaries (human- vs. machine-generated.) An ANOVA revealed a reliable effect of summary type: $F(1; 15) = 20.38$, $p < .01$ indicating that human summaries are perceived as significantly more coherent than system-generated ones. Finally, we also compared the elicited ratings against the DUC evaluations using correlation analysis. The human judgments were discretized to two classes (i.e., 0 or 1) using entropy-based discretization (Witten and Frank 2000). We found a moderate correlation between the human ratings and DUC coherence violations ($r = .41$, $p < .01$). This is expected given that DUC evaluators were using a different scale and and were not explicitly assessing summary coherence.

The summaries used in our rating elicitation study form the basis of a corpus used for the development of our entity-based coherence models. To increase the size of our training and test sets, we augmented the materials used in the elicitation study with additional DUC summaries generated by humans for the same input sets. We assumed that these summaries were maximally coherent. As mentioned previously, our participants tend to rate human-authored summaries higher than machine-generated ones. 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 6×16 summaries (average length 4.8), yielding $\binom{6}{2} \times 16 = 240$ pairwise rankings. Because human summaries often have identical (high) scores, we eliminated pairs of such summaries from the training set. Consequently, the resulting training corpus consisted of 144 summaries. In a similar fashion, we obtained 80 pairwise rankings for the test set. Six documents from the training data were used as a development set.

Features, Parameter Settings, and Training Requirements. We examine the influence of linguistic knowledge on model performance by comparing models with varying degrees of linguistic complexity. To be able to assess the performance of our models across tasks (e.g., sentence ordering vs. summarization), we experimented with the same model types introduced in the previous experiment (see Section 4.3). We also investigate the training requirements for these models on the summary coherence task.

Experiment 1 differs from the present study in the way coreference information was obtained. In Experiment 1, a coreference resolution tool was applied to human-written texts, which are grammatical and coherent. Here, we apply a coreference tool to automatically generated summaries. Because many summaries in our corpus are fraught with coherence violations, the performance of a coreference resolution tool is likely to drop. Unfortunately, resolving coreference in the input documents would require a multi-document coreference tool, which is currently unavailable to us.

As in Experiment 1, the frequency threshold and the length of the transition sequence were optimized on the development set. Optimal salience-based models were obtained for entities with frequency ≥ 2. The optimal transition length was ≤ 2. All models were trained and tested using SVM^{light} (Joachims 2002).

Comparison with State-of-the-Art Methods. Our results were compared to the LSA model introduced in Experiment 1 (see Section 4.2 for details). Unfortunately, we could not

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

employ Barzilay and Lee's (2004) content models for the summary ranking task. Being domain-dependent, these models require access to domain representative texts for training. Our summary corpus, however, contains texts from multiple domains and does not provide an appropriate sample for reliably training content models.

5.3 Results

Impact of Linguistic Representation. Our results are summarized in Table 8. Similarly to the sentence ordering task, we observe that the linguistically impoverished model **Coreference**−**Syntax**−**Salience**− exhibits decreased accuracy when compared against models that operate over more sophisticated representations. However, the contribution of individual knowledge sources differs in this task. For instance, coreference resolution improved model performance in ordering, but it causes a decrease in accuracy in summary evaluation (compare the models **Coreference**+**Syntax**+**Salience**+ and **Coreference**−**Syntax**+**Salience**+ in Tables 5 and 8). This drop in performance can be attributed to two factors both related to the fact that our summary corpus contains many machine-generated texts. First, an automatic coreference resolution tool will be expected to be less accurate on our corpus, 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.

Both salience and syntactic information contribute to the accuracy of the ranking model. The impact of each of these knowledge sources in isolation is not dramatic—dropping either of them yields some decrease in accuracy, but the difference is not statistically significant. However, eliminating both salience and syntactic information significantly decreases performance (compare **Coreference**−**Syntax**+**Salience**+ against **Coreference**+**Syntax**−**Salience**− and **Coreference**−**Syntax**−**Salience**− in Table 8).

Figure 2 shows the learning curve for our best model **Coreference**−**Syntax**+**Salience**+. Although the model performs poorly when trained on a small fraction of the data, it stabilizes relatively fast (with 80 pairwise rankings), and does not improve after

Table 8
Summary ranking accuracy measured as fraction of correct pairwise rankings in the test set. **Coreference[+/−]** indicates whether anaphoric information has been used when constructing the entity grid. Similarly, **Syntax[+/−]** and **Salience[+/−]** reflect the use of syntactic and salience information. Diacritics ** ($p < .01$) and * ($p < .05$) indicate whether **Coreference**−**Syntax**+**Salience**+ is significantly different from all other models (using a Fisher Sign Test).

Model	Accuracy
Coreference+Syntax+Salience+	80.0
Coreference+Syntax+Salience−	75.0
Coreference+Syntax−Salience+	78.8
Coreference−**Syntax**+**Salience**+	**83.8**
Coreference+Syntax−Salience−	71.3*
Coreference−Syntax+Salience−	78.8
Coreference−Syntax−Salience+	77.5
Coreference−Syntax−Salience−	73.8*
Latent Semantic Analysis	52.5**

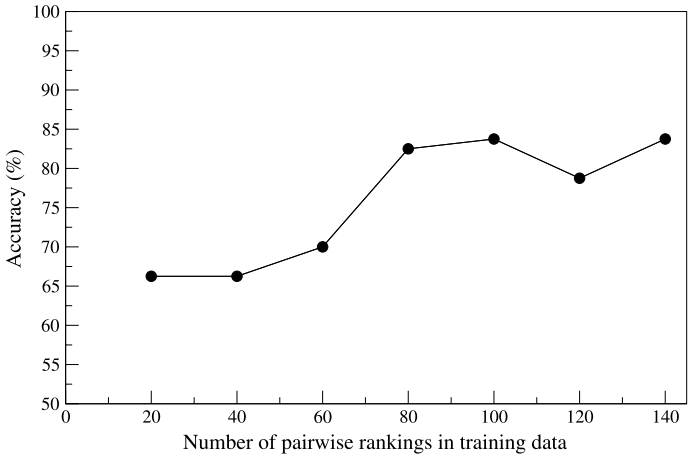

Figure 2
Learning curve for the entity-based model **Coreference**−**Syntax**+**Salience**+ applied to the summary ranking task.

a certain point. These results suggest that further improvements to summary ranking are unlikely to come from adding more annotated data.

Comparison with the State-of-the-Art. As in Experiment 1, we compared the best performing grid model (**Coreference**−**Syntax**+**Salience**+) against LSA (see Table 8). The former model significantly outperforms the latter ($p < .01$) by a wide margin. LSA is perhaps at a disadvantage here because it has been exposed only to human-authored texts. Machine-generated summaries are markedly distinct from human texts even when these are incoherent (as in the case of our ordering experiment). For example, 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.

Furthermore, note that in contrast to the documents in Experiment 1, the summaries being ranked here differ in lexical choice. Some are written by humans (and are thus abstracts), whereas others have been produced by systems following different summarization paradigms (some systems perform rewriting whereas others extract sentences verbatim from the source documents). This means that LSA may consider a summary coherent simply because its vocabulary is familiar (i.e., it contains words for which reliable vectors have been obtained). Analogously, a summary with a large number of out-of-vocabulary lexical items will be given low similarity scores, irrespective of whether it is coherent or not. This is not uncommon in summaries with many proper names. These often do not overlap with the proper names found in the North American News Corpus used for training the LSA model. Lexical differences exert much less influence on the entity-grid model which abstracts away from alternative verbalizations of the same content and captures coherence solely on the basis of grid topology.

6. Experiment 3: Readability Assessment

So far, our experiments have explored the potential of the proposed discourse representation for coherence modeling. We have presented several classes of grid models

achieving good performance in discerning coherent from incoherent texts. Our experiments also reveal a surprising property of grid models: Even though these models are not lexicalized, they are domain- and style-dependent. In this section, we investigate in detail this feature of grid models. Here, we move away from the coherence rating task and put the entity-grid representation further to the test by examining whether it can be usefully employed in style classification. Specifically, we embed our entity grids into a system that assesses document readability. The term describes the ease with which a document can be read and understood. The quantitative measurement of readability has attracted considerable interest and debate over the last 70 years (see Mitchell [1985] and Chall [1958] for detailed overviews) and has recently benefited from the use of NLP technology (Schwarm and Ostendorf 2005).

A number of readability formulas have been developed with the primary aim of assessing whether texts or books are suitable for students at particular grade levels or ages. Many readability methods focus on simple approximations of *semantic* factors concerning the words used and *syntactic* factors concerning the length or structure of sentences (Gunning 1952; Kincaid et al. 1975; Chall and Dale 1995; Stenner 1996; Katz and Bauer 2001). Despite their widespread applicability in education and technical writing (Kincaid et al. 1981), readability formulas are often criticized for being too simplistic; they systematically ignore many important factors that affect readability such as discourse coherence and cohesion, layout and formatting, use of illustrations, the nature of the topic, the characteristics of the readers, and so forth.

Schwarm and Ostendorf (2005) developed a method for assessing readability which addresses some of the shortcomings of previous approaches. By recasting readability assessment as a classification task, they are able to combine several knowledge sources ranging from traditional reading level measures, to statistical language models, and syntactic analysis. Evaluation results show that their system outperforms two commonly used reading level measures (the Flesch-Kincaid Grade Level index and Lexile). In the following we build on their approach and examine whether the entity-grid representation introduced in this article contributes to the readability assessment task. The incorporation of coherence-based information in the measurement of text readability is, to our knowledge, novel.

6.1 Modeling

We follow Schwarm and Ostendorf (2005) in treating readability assessment as a classification task. The unit of classification is a single article and the learner's task is to predict whether it is easy or difficult to read. A variety of machine learning techniques are amenable to this problem. Because our goal was to replicate Schwarm and Ostendorf's system as closely as possible, we followed their choice of support vector machines (SVMs) (Joachims 1998b) for our classification experiments. Our training sample therefore consisted of n documents such that

$$(\vec{x_1}, y_1), \ldots, (\vec{x_n}, y_n) \qquad \vec{x_i} \in \Re^N, y_i \in \{-1, +1\}$$

where $\vec{x_i}$ is a feature vector for the ith document in the training sample and y_i its (positive or negative) class label. In the basic SVM framework, we try to separate the positive and negative instances by a hyperplane. This means that there is a weight

Table 9
Excerpts from the Britannica readability corpus

The Lemma Valletta in *Britannica*
Also spelled Valletta, seaport and capital of Malta, on the northeast coast of the island. The nucleus of the city is built on the promontory of Mount Sceberras that runs like a tongue into the middle of a bay, which it thus divides into two harbours, Grand Harbour to the east and Marsamxett (Marsamuscetto) Harbour to the west. Built after the Great Siege of Malta in 1565, which checked the advance of Ottoman power in southern Europe, it was named after Jean Parisot de la Valette, grand master of the order of Hospitallers (Knights of St. John of Jerusalem), and became the Maltese capital in 1570. The Hospitallers were driven out by the French in 1798, and a Maltese revolt against the French garrison led to Valletta's seizure by the British in 1800.
The Lemma Valletta in *Britannica Elementary*
A port city, Valletta is the capital of the island country of Malta in the Mediterranean Sea. Valletta is located on the eastern coast of the largest island, which is also named Malta. Valletta lies on a peninsula—a land mass surrounded by water on three sides. It borders Marsamxett Harbor to the north and Grand Harbor to the south. The eastern end of the city juts out into the Mediterranean. Valletta was planned in the 16th century by the Italian architect Francesco Laparelli. To make traveling through Valletta easier, Laparelli designed the city in a grid pattern with straight streets that crossed each other and ran the entire width and length of the town. Valletta was one of the first towns to be laid out in this way.

vector $\mathbf{w}$ and a threshold b, so that all positive training examples are on one side of the hyperplane, while all negative ones lie on the other side. This is equivalent to requiring

$$y_i[(\mathbf{w} \cdot \vec{x_i}) + b] > 0$$

Finding the optimal hyperplane is an optimization problem which can be solved efficiently using the procedure described in Vapnik (1998). SVMs have been widely used for many NLP tasks ranging from text classification (Joachims 1998b), to syntactic chunking (Kudo and Matsumoto 2001), and shallow semantic parsing (Pradhan et al. 2005).

6.2 Method

Data. For our experiments we used a corpus collected by Barzilay and Elhadad (2003) from the *Encyclopedia Britannica* and *Britannica Elementary*. The latter is a new version targeted at children. The corpus contains 107 articles from the full version of the encyclopedia and their corresponding simplified articles from *Britannica Elementary* (214 articles in total). Although these texts are not explicitly annotated with grade levels, they still represent two broad readability categories, namely, easy and difficult.[11] Examples of these two categories are given in Table 9.

11 The Britannica corpus was also used by Schwarm and Ostendorf (2005); in addition they make use of a corpus compiled from the *Weekly Reader*, an educational newspaper with documents targeted at grade levels 2–5. Unfortunately, this corpus is not publicly available.

Features and Parameter Settings. We created two system versions: the first one used solely Schwarm and Ostendorf (2005) features;[12] the second one employed a richer feature space—we added the entity-based representation proposed here to their original feature set. We will briefly describe the readability-related features used in our systems and direct the interested reader to Schwarm and Ostendorf for a more detailed discussion.

Schwarm and Ostendorf (2005) use three broad classes of features: syntactic, semantic, and their combination. Their syntactic features are average sentence length and features extracted from parse trees computed using Charniak's (2000) parser. The latter include average parse tree height, average number of NPs, average number of VPs, and average number of subordinate clauses (SBARs). We computed average sentence length by measuring the number of tokens per sentence.

Their semantic features include the average number of syllables per word, and language model perplexity scores. A unigram, bigram, and trigram model was estimated for each class, and perplexity scores were used to assess their performance on test data. Following Schwarm and Ostendorf (2005) we used information gain to select words that were good class discriminants. All remaining words were replaced by their parts of speech. The vocabulary thus consisted of 300 words with high information gain and 36 Penn Treebank part-of-speech tags. The language models were estimated using maximum likelihood estimation and smoothed with Witten-Bell discounting. The language models described in this article were all built using the CMU statistical language modeling toolkit (Clarkson and Rosenfeld 1997). Our perplexity scores were six in total (2 classes $\times$ 3 language models).

Finally, the Flesch-Kincaid Grade Level score was included as a feature that captures both syntactic and semantic text properties. The Flesch-Kincaid formula estimates readability as a combination of the the average number of syllables per word and the average number of words per sentence:

$$0.39\left(\frac{\textit{total words}}{\textit{total sentences}}\right) + 11.8\left(\frac{\textit{total syllables}}{\textit{total words}}\right) - 15.59 \tag{3}$$

We also enriched Schwarm and Ostendorf's (2005) feature space with coherence-based features. Each document was represented as a feature vector using the entity transition notation introduced in Section 3. We experimented with two models that yielded good performances in our previous experiments: **Coreference**+**Syntax**+**Salience**+ (see Experiment 1) and **Coreference**−**Syntax**+**Salience**+ (see Experiment 2). The transition length was $\leq$2 and entities were considered salient if they occurred $\geq$2 times. As in our previous experiments, we compared the entity-based representation against LSA. The latter is a measure of the semantic relatedness across pairs of sentences. We could not apply the HMM-based content models (Barzilay and Lee 2004) to the readability data set. The encyclopedia lemmas are written by different authors and consequently vary considerably in structure and vocabulary choice. Recall that these models are suitable for more restricted domains and texts that are more formulaic in nature.

12 Schwarm and Ostendorf (2005) define out-of-vocabulary (OOV) scores relative to the most common words in grade 2, the lowest grade level in their corpus; it was not possible to estimate OOV scores, because we did not have access to grade 2 texts.

Table 10
The contribution of coherence-based features to the automatic readability assessment task. Diacritics ** ($p < .01$) and * ($p < .05$) indicate whether differences in accuracy between Schwarm and Ostendorf and all other models are significant (using a Fisher Sign test).

Model	Accuracy
Schwarm & Ostendorf	78.56
Schwarm & Ostendorf, Coreference+Syntax+Salience+	**88.79***
Schwarm & Ostendorf, Coreference−Syntax+Salience+	79.49
Schwarm & Ostendorf, Latent Semantic Analysis	78.56
Coreference+Syntax+Salience+	50.90**
Coreference−Syntax+Salience+	49.55**
Latent Semantic Analysis	48.58**

The different systems were trained and tested on the Britannica corpus using five-fold cross-validation.[13] The language models were created anew for every fold using the documents in the training data. We use Joachims' (1998a) SVM^{light} package for training and testing with all parameters set to their default values.

Evaluation Metric. We measure classification accuracy (i.e., the number of classes assigned correctly by the SVM over the size of the test set). We report accuracy averaged over folds. A chance baseline (selecting one class at random) yields an accuracy of 50%. Our training and test sets have the same number of documents for the two readability categories.

6.3 Results

Table 10 summarizes our results on the readability assessment task. We first compared Schwarm and Ostendorf's (2005) system against a system that incorporates entity-based coherence features (see rows 3–4 in Table 10). As can be seen, the system's accuracy significantly increases by 10% when the full feature set is included (**Coreference**+**Syntax**+**Salience**+). Entity-grid features that do not incorporate coreference information (**Coreference**−**Syntax**+**Salience**+) perform numerically better (compare row 1 and 3 in Table 10); however, the difference is not statistically significant.

The superior performance of the **Coreference**+**Syntax**+**Salience**+ feature set is not entirely unexpected. Inspection of our corpus revealed that easy and difficult texts differ in their distribution of pronouns and coreference chains in general. Easy texts tend to employ less coreference and the use of personal pronouns is relatively sparse. To give a concrete example, the pronoun *they* is attested 173 times in the difficult corpus and only 73 in the easy corpus. This observation suggests that coreference information is a good indicator of the level of reading difficulty and explains why its omission from the entity-based feature space yields inferior performance.

13 The data for the experiments reported here can be found at `http://homepages.inf.ed.ac.uk/mlap/coherence/`.

Furthermore, note that discourse-level information is absent from Schwarm and Ostendorf's (2005) original model. The latter employs a large number of lexical and syntactic features which capture sentential differences among documents. Our entity-based representation supplements their feature space with information spanning two or more successive sentences. We thus are able to model stylistic differences in readability that go beyond syntax and lexical choice. Besides coreference, our feature representation captures important information about the presence and distribution of entities in discourse. For example, difficult texts tend to have twice as many entities as easy ones. Consequently, easy and difficult texts are represented by entity transition sequences with different probabilities (e.g., the sequences **[S S]** and **[S O]** are more probable in difficult texts). Interestingly, when coherence is quantified using LSA, we observe no improvement to the classification task. The LSA scores capture lexical or semantic text properties similar to those expressed by the Flesch Kincaid index and the perplexity scores (e.g., word repetition). It is therefore not surprising that their inclusion in the feature set does not increase performance.

We also evaluated the training requirements for the readability system described herein. Figure 3 shows the learning curve for Schwarm and Ostendorf's (2005) model enhanced with the **Coreference**+**Syntax**+**Salience**+ feature space and on its own. As can be seen, both models perform relatively well when trained on small data sets (e.g., 20–40 documents) and reach peak accuracy with half of the training data. The inclusion of discourse-based features consistently increases accuracy irrespective of the amount of training data available. Figure 3 thus suggests that better feature engineering is likely to bring further performance improvements on the readability task.

Our results indicate that the entity-based text representation introduced here captures aspects of text readability and can be successfully incorporated into a practical system. Coherence is by no means the sole predictor of readability. In fact, on its own, it performs poorly on this task as demonstrated when using either LSA or the entity-based feature space without Schwarm and Ostendorf's (2005) features (see rows 5–7 in Table 10). Rather, we claim that coherence is one among many factors contributing to text readability and that our entity-grid representation is well-suited for text classification tasks such as reading level assessment.

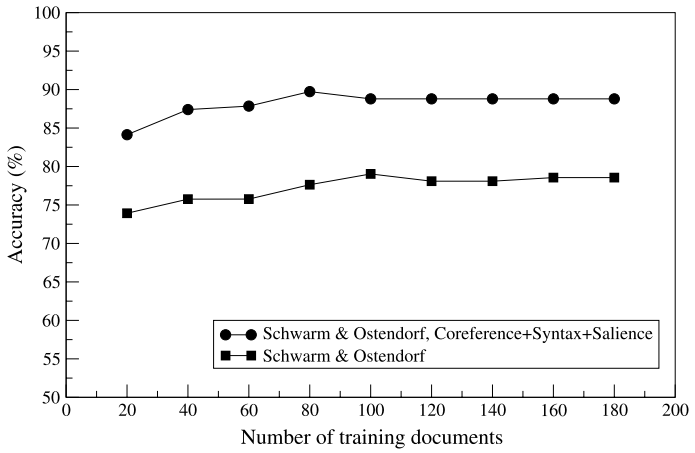

Figure 3
Learning curve for Schwarm and Ostendorf's (2005) model on its own and enhanced with the **Coreference**+**Syntax**+**Salience**+ feature space.

7. Discussion and Conclusions

In this article 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 learning task and show that our entity-based representation is well-suited for ranking-based generation and text classification tasks. Using the proposed representation, we achieve good performance on text ordering, summary coherence evaluation, and readability assessment.

The entity grid is a flexible, yet computationally tractable, representation. We investigated three important parameters for grid construction: the computation of coreferring entity classes, the inclusion of syntactic knowledge, and the influence of salience. All these knowledge sources figure prominently in theories of discourse (see Section 2) and are considered important in determining coherence. Our results empirically validate the importance of salience and syntactic information (expressed by **S**, **O**, **X**, and **–**) for coherence-based models. The combination of both knowledge sources (Syntax+Salience) yields models with consistently good performance for all our tasks.

The benefits of full coreference resolution are less uniform. This is partly due to mismatches between training and testing conditions. The system we employ (Ng and Cardie 2002) was trained on human-authored newspaper texts. The corpora we used in our sentence ordering and readability assessment experiments are somewhat similar (i.e., human-authored narratives), whereas our summary coherence rating experiment employed machine generated texts. It is therefore not surprising that coreference resolution delivers performance gains on the first two tasks but not on the latter (see Table 5 in Section 4 and Table 10 in Section 6.3). Our results further show that in lieu of an automatic coreference resolution system, entity classes can be approximated simply by string matching. The latter is a good indicator of nominal coreference; it is often included as a feature in machine learning approaches to coreference resolution (Soon, Ng, and Lim 2001; Ng and Cardie 2002) and is relatively robust (i.e., likely to deliver consistent results in the face of different domains and genres).

It is important to note that, although inspired by entity-based theories of discourse coherence, our approach is not a direct implementation of any theory in particular. Rather, we sacrifice linguistic faithfulness for automatic computation and breadth of coverage. Despite approximations and unavoidable errors (e.g., in the parser's output), our results indicate that entity grids are a useful representational framework across tasks and text genres. In agreement with Poesio et al. (2004) we find that pronominalization is a good indicator of document coherence. We also find that coherent texts are characterized by transitions with particular properties which do not hold for all discourses. Contrary to Centering Theory, we remain agnostic to the type of transitions that our models capture (e.g., CONTINUE, SHIFT). We simply record whether an entity is mentioned in the discourse and in what grammatical role. Our experiments quantitatively measured the predictive power of various linguistic features for several coherence-related tasks. Crucially, we find that our models are sensitive to the domain at hand and the type of texts under consideration (human-authored vs. machine generated texts). This is an unavoidable consequence of the grid representation, which is entity-specific. Differences in entity distribution indicate not only differences in coherence, but also in writing conventions and style. Similar observations have been made in other work which is closer in spirit to Centering's claims (Hasler 2004; Karamanis et al. 2004; Poesio et al. 2004).

An important future direction lies in augmenting our entity-based representation with more fine-grained 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 De Mori 1990) seem particularly promising in this context. The granularity of syntactic information is another topic that warrants further investigation. So far we have only considered the contribution of "core" grammatical relations to the grid construction. Expanding our grammatical categories to modifiers and adjuncts may provide additional information, in particular when considering machine generated texts. We also plan to investigate whether the proposed discourse representation and modeling approaches generalize across different languages. For instance the identification and extraction of entities poses additional challenges in grid construction for Chinese where word boundaries are not denoted orthographically (by space). Similar challenges arise in German, a language with a large number of inflected forms and productive derivational processes (e.g., compounding) not indicated by orthography.

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 coreference system on our data. Thanks to Eli Barzilay, Eugene Charniak, Michael Elhadad, Noemie Elhadad, Nikiforos Karamanis, 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 herein are those of the authors and do not necessarily reflect the views of the National Science Foundation or EPSRC.

References

Althaus, Ernst, Nikiforos Karamanis, and Alexander Koller. 2004. Computing locally coherent discourses. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 399–406, Barcelona, Spain.

Ariel, Mira. 1988. Referring and accessibility. *Journal of Linguistics*, 24:65–87.

Asher, Nicholas and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge, England.

Barzilay, Regina. 2003. *Information Fusion for Multi-Document Summarization: Praphrasing and Generation*. Ph.D. thesis, Columbia University, New York.

Barzilay, Regina and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 8th Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Sapporo, Japan.

Barzilay, Regina and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 141–148, Ann Arbor, MI.

Barzilay, Regina and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of the 2nd Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 113–120, Boston, MA.

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

Brennan, Susan E., Marilyn W. Friedman, and Charles J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the*

25th Annual Meeting of the Association for Computational Linguistics, pages 155–162, Palo Alto, CA.

Briscoe, Ted and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1499–1504, Las Palmas, Canary Islands.

Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li, editor, *Subject and Topic*. Academic Press, New York, pages 25–55.

Chall, Jeanne S. 1958. *Readability: An Appraisal of Research and Application*. Number 34 in Bureau of Educational Research Monographs. Ohio State University Press, Columbus.

Chall, Jeanne S. and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Cambridge, MA.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 132–139, Seattle, WA.

Clark, Herbert H. and Susan E. Haviland. 1977. Comprehension and the given-new contract. In Roy O. Freedle, editor, *Discourse Production and Comprehension*. Ablex, Norwood, NJ, pages 1–39.

Clarkson, Philip and Ronald Rosenfeld. 1997. Statistical language modeling. In *Proceedings of ESCA EuroSpeech'97*, pages 2707–2710, Rhodes, Greece.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23, Madrid, Spain.

Collins, Michael. 2002. Discriminative reranking for natural language parsing. In *Proceedings of the 17th International Conference on Machine Learning*, pages 175–182, Palo Alto, CA.

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

Freund, Yovav, Raj Iyer, Robert E. Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferencs. *Machine Learning*, 4:933–969.

Germann, Ulrich, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. 2004. Fast and optimal decoding for machine translation. *Artificial Intelligence*, 154(1–2):127–143.

Givon, Talmy. 1987. Beyond foreground and background. In Russell S. Tomlin, editor, *Coherence and Grounding in Discourse*. Benjamins, Amsterdam/Philadelphia, pages 175–188.

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

Gundel, Jaenette K., Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.

Gunning, Robert. 1952. *The Technique of Clear Writing*. McGraw Hill, New York.

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

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

Hoey, Michael. 1991. *Patterns of Lexis in Text*. Oxford University Press, Oxford, England.

Hudson, S. B., M. K. Tanenhaus, and G. S. Dell. 1986. The effect of the discourse center on the local coherence of a discourse. In *Proceedings of the 8th Annual Meeting of the Cognitive Science Society*, pages 96–101, Amherst, MA.

Joachims, Thorsten. 1998a. Making large-scale support vector machine learning practical. In Bernard Schölkopf, Christopher Burges, and Alexander Smola, editors, *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA.

Joachims, Thorsten. 1998b. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142, Berlin, Springer.

Joachims, Thorsten. 2002. Optimizing search engines using clickthrough data. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining*, pages 133–142, Chicago, IL.

Kameyama, Megumi. 1986. A property-sharing constraint in centering. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 200–206, New York.

Karamanis, Nikiforos. 2003. *Entity Coherence for Descriptive Text Structuring*. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

Karamanis, Nikiforos, Massimo Poesio, Chris Mellish, and Jon Oberlander. 2004. Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 391–398, Barcelona, Spain.

Karttunen, Lauri. 1976. Discourse referents. In James D. McCawley, editor, *Syntax and Semantics: Notes from the Linguistic Underground*, volume 7. Academic Press, New York, pages 363–386.

Katz, Irvin R. and Malcolm I. Bauer. 2001. Sourcefinder: Course preparation via linguistically targeted web search. *Journal of Educational Technology and Society*, 4(3):45–49.

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

Kincaid, J. Peter, James Aagard, John O'Hara, and Larry Cottrell. 1981. Computer readability editing system. *IEEE Transactions on Professional Communication*, 1(24):34–81.

Kincaid, Peter J., Robert P. Fishburne, Richard L. Rodgers, and Brad S. Chissom. 1975. Derivation of new readability formulas for Navy enlisted personnel. Research Branch Report 8-75, U.S. Naval Air Station, Memphis, TN.

Knight, Kevin and Vasileios Hatzivassiloglou. 1995. Two-level, many-path generation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 252–260, Cambridge, MA.

Kudo, Taku and Yuji Matsumoto. 2001. Chunking with support vector machines. In Thorsten Joachims, editor, *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 192–199, Pittsburgh, PA.

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

Kuno, Susumu. 1972. Functional sentence perspective. *Linguistic Inquiry*, 3:269–320.

Landauer, Thomas K. and Susan 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.

Langkilde, Irene and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, pages 704–710, Montréal, Canada.

Lapata, Mirella. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 545–552, Sapporo, Japan.

Lin, Chin-Yew and Eduard H. Hovy. 2003. Automatic evaluation of summaries using *n*-gram co-occurrence statistics. In *Proceedings of the 2nd Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 71–78, Boston, MA.

Lin, Dekang. 2001. LaTaT: Language and text analysis tools. In *Proceedings of the 1st International Conference on Human Language Technology Research*, pages 222–227, San Francisco, CA.

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

Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA.

McKoon, Gail and Roger Ratcliff. 1992. Inference during reading. *Psychological Review*, 99(3):440–446.

Mellish, Chris, Mick O'Donnell, Jon Oberlander, and Alistair Knott. 1998. Experiments using stochastic search for text planning. In *Proceedings of the 9th International Workshop on Natural Language Generation*, pages 98–107, New Brunswick, NJ.

Miltsakaki, Eleni and Karen Kukich. 2000. The role of centering theory's rough-shift in the teaching and evaluation of writing skills. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 408–415, Hong Kong.

Mitchell, James V. 1985. *The Ninth Mental Measurements Yearbook*. University of Nebraska Press, Lincoln.

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

Ng, Vincent and Claire Cardie. 2002. Improving machine learning approaches

to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, PA.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Poesio, Massimo, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–363.

Pradhan, Sameer, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, and Dan Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 60(1):11–39.

Prince, Ellen. 1978. A comparison of *wh*-clefts and *it*-clefts in discourse. *Language*, 54:883–906.

Prince, Ellen. 1981. Toward a taxonomy of given-new information. In Peter Cole, editor, *Radical Pragmatics*. Academic Press, New York, pages 223–255.

Reiter, Ehud and Robert Dale. 2000. *Building Natural-Language Generation Systems*. Cambridge University Press, Cambridge, England.

Schwarm, Sarah E. and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 523–530, Ann Arbor, MI.

Scott, Donia and Clarisse Sieckenius de Souza. 1990. Getting the message across in RST-based text generation. In Robert Dale, Chris Mellish, and Michael Zock, editors, *Current Research in Natural Language Generation*. Academic Press, New York, pages 47–73.

Sidner, Candace L. 1979. *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. Ph.D. thesis, MIT.

Soon, W. M., Hwee Tou Ng, and D. C. Y. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Stenner, A. Jackson. 1996. Measuring reading comprehension with the lexile framework. Presented at the California Comparability Symposium, Burlingame, CA.

Strube, Michael and Udo Hahn. 1999. Functional centering—Grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309–344.

Toutanova, Kristina, Penka Markova, and Christopher D. Manning. 2004. The leaf projection path view of parse trees: Exploring string kernels for HPSG parse selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 166–173, Barcelona, Spain.

Vapnik, Vladimir. 1998. *Statistical Learning Theory*. Wiley, Chichester, UK.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45–52, San Francisco, CA.

Walker, Marilyn, Masayo Iida, and Sharon Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–232.

Walker, Marilyn, Aravind Joshi, and Ellen Prince, editors. 1998. *Centering Theory in Discourse*. Clarendon Press, Oxford, UK.

Walker, Marilyn A., Owen Rambow, and Monica Rogati. 2001. Spot: A trainable sentence planner. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 17–24, Pittsburgh, PA.

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

Witten, Ian H. and Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufman, San Mateo, CA.