# A Unified Local and Global Model for Discourse Coherence

**Micha Elsner, Joseph Austerweil, and Eugene Charniak**
Brown Laboratory for Linguistic Information Processing (BLLIP)
Brown University
Providence, RI 02912
{melsner,ec}@cs.brown.edu, joseph.austerweil@gmail.com

## Abstract

We present a model for discourse coherence which combines the local entity-based approach of (Barzilay and Lapata, 2005) and the HMM-based content model of (Barzilay and Lee, 2004). Unlike the mixture model of (Soricut and Marcu, 2006), we learn local and global features jointly, providing a better theoretical explanation of how they are useful. As the local component of our model we adapt (Barzilay and Lapata, 2005) by relaxing independence assumptions so that it is effective when estimated generatively. Our model performs the ordering task competitively with (Soricut and Marcu, 2006), and significantly better than either of the models it is based on.

## 1 Introduction

Models of coherent discourse are central to several tasks in natural language processing: such models have been used in text generation (Kibble and Power, 2004) and evaluation of human-produced text in educational applications (Miltsakaki and Kukich, 2004; Higgins et al., 2004). Moreover, an accurate model can reveal information about document structure, aiding in such tasks as supervised summarization (Barzilay and Lapata, 2005).

Models of coherence tend to fall into two classes. Local models (Lapata, 2003; Barzilay and Lapata, 2005; Foltz et al., 1998) attempt to capture the generalization that adjacent sentences often have similar content, and therefore tend to contain related words. Models of this type are good at finding sentences that belong near one another in the document. However, they have trouble finding the beginning or end of the document, or recovering from sudden shifts in topic (such as occur at paragraph boundaries). Some local models also have trouble deciding which of a pair of related sentences ought to come first.

In contrast, the global HMM model of Barzilay and Lee (2004) tries to track the predictable changes in topic between sentences. This gives it a pronounced advantage in ordering sentences, since it can learn to represent beginnings, ends and boundaries as separate states. However, it has no local features; the particular words in each sentence are generated based only on the current state of the document. Since information can pass from sentence to sentence only in this restricted manner, the model sometimes fails to place sentences next to the correct neighbors.

We attempt here to unify the two approaches by constructing a model with both sentence-to-sentence dependencies providing local cues, and a hidden topic variable for global structure. Our local features are based on the entity grid model of (Barzilay and Lapata, 2005; Lapata and Barzilay, 2005). This model has previously been most successful in a conditional setting; to integrate it into our model, we first relax its independence assumptions to improve its performance when used generatively. Our global model is an HMM like that of Barzilay and Lee (2004), but with emission probabilities drawn from the entity grid. We present results for two tasks, the ordering task, on which global models usually do well, and the discrimination task, on which local models tend to outperform them. Our model improves on purely global or local approaches on both


*Proceedings of NAACL HLT 2007*, pages 436–443,
Rochester, NY, April 2007. 

tasks.

Previous work by Soricut and Marcu (2006) has also attempted to integrate local and global features using a mixture model, with promising results. However, mixture models lack explanatory power; since each of the individual component models is known to be flawed, it is difficult to say that the combination is theoretically more sound than the parts, even if it usually works better. Moreover, since the model we describe uses a strict subset of the features used in the component models of (Soricut and Marcu, 2006), we suspect that adding it to the mixture would lead to still further improved results.

## 2 Naive Entity Grids

Entity grids, first described in (Lapata and Barzilay, 2005), are designed to capture some ideas of Centering Theory (Grosz et al., 1995), namely that adjacent utterances in a locally coherent discourses are likely to contain the same nouns, and that important nouns often appear in syntactically important roles such as subject or object. An entity grid represents a document as a matrix with a column for each entity, and a row for each sentence. The entry $r_{i,j}$ describes the syntactic role of entity $j$ in sentence $i$: these roles are subject (**S**), object (**O**), or some other role (**X**)[1]. In addition there is a special marker (-) for nouns which do not appear at all in a given sentence. Each noun appears only once in a given row of the grid; if a noun appears multiple times, its grid symbol describes the most important of its syntactic roles: subject if possible, then object, or finally other. An example text is figure 1, whose grid is figure 2.

Nouns are also treated as salient or non-salient, another important concern of Centering Theory. We condition events involving a noun on the frequency of that noun. Unfortunately, this way of representing salience makes our model slightly deficient, since the model conditions on a particular noun occurring e.g. 2 times, but assigns nonzero probabilities to documents where it occurs 3 times. This is theoretically quite unpleasant but in comparing different orderings of the same document, it seems not to do too much damage.

0 [The commercial pilot]$_{\mathbf{O}}$ , [sole occupant of [the airplane]$_{\mathbf{X}}$]$_{\mathbf{X}}$ , was not injured .
1 [The airplane]$_{\mathbf{O}}$ was owned and operated by [a private owner]$_{\mathbf{X}}$ .
2 [Visual meteorological conditions]$_{\mathbf{S}}$ prevailed for [the personal cross country flight for which [a VFR flight plan]$_{\mathbf{O}}$ was filed]$_{\mathbf{X}}$ .
3 [The flight]$_{\mathbf{S}}$ originated at [Nuevo Laredo , Mexico]$_{\mathbf{X}}$ , at [approximately 1300]$_{\mathbf{X}}$.

Figure 1: A section of a document, with syntactic roles of noun phrases marked.

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| PLAN | - | - | O | - |
| AIRPLANE | X | O | - | - |
| CONDITION | - | - | S | - |
| FLIGHT | - | - | X | S |
| PILOT | O | - | - | - |
| PROPER | - | - | - | X |
| OWNER | - | X | - | - |
| OCCUPANT | X | - | - | - |

Figure 2: The entity grid for figure 1[2].

Properly speaking entities may be referents of many different nouns and pronouns throughout the discourse, and both (Lapata and Barzilay, 2005) and (Barzilay and Lapata, 2005) present models which use coreference resolution systems to group nouns. We follow (Soricut and Marcu, 2006) in dropping this component of the system, and treat each head noun as having an individual single referent.

To model transitions in this entity grid model, Lapata and Barzilay (2005) takes a generative approach. First, the probability of a document is defined as $P(D) = P(S_i..S_n)$, the joint probability of all the sentences. Sentences are generated in order conditioned on all previous sentences:

$$P(D) = \prod_i P(S_i|S_{0..(i-1)}). \quad (1)$$

We make a Markov assumption of order $h$ (in our experiments $h = 2$) to shorten the history. We represent the truncated history as $\vec{S}^h_{i-1} = S_{(i-h)}..S_{(i-1)}$.

Each sentence $S_i$ can be split up into a set of nouns representing entities, $E_i$, and their corresponding syntactic roles $R_i$, plus a set of words which are not entities, $W_i$. The model treats $W_i$ as independent of the previous sentences. For any fixed

[1] Roles are determined heuristically using trees produced by the parser of (Charniak and Johnson, 2005). Following previous work, we slightly conflate thematic and syntactic roles, marking the subject of a passive verb as **O**.

[2] The numeric token "1300" is removed in preprocessing, and "Nuevo Laredo" is marked as "PROPER".

set of sentences $S_i$, $\prod_i P(W_i)$ is always constant, and so cannot help in finding a coherent ordering. The probability of a sentence is therefore dependent only on the entities:

$$P(S_i|\vec{S}^h_{(i-1)}) = P(E_i, R_i|\vec{S}^h_{(i-1)}). \quad (2)$$

Next, the model assumes that each entity $e_j$ appears in sentences and takes on syntactic roles independent of all the other entities. As we show in section 3, this assumption can be problematic. Once we assume this, however, we can simplify $P(E_i, R_i|\vec{S}^h_{(i-1)})$ by calculating for each entity whether it occurs in sentence $i$ and if so, which role it takes. This is equivalent to predicting $r_{i,j}$. We represent the history of the specific entity $e_j$ as $\vec{r}^h_{(i-1),j} = r_{(i-h),j}..r_{(i-1),j}$, and write:

$$P(E_i, R_i|\vec{S}^h_{(i-1)}) \approx \prod_j P(r_{i,j}|\vec{r}^h_{(i-1),j}). \quad (3)$$

For instance, in figure 2, the probability of $S_3$ with horizon 1 is the product of $P(\mathbf{S}|\mathbf{X})$ (for FLIGHT), $P(\mathbf{X}|-)$ (for PROPER), and likewise for each other entity, $P(-|\mathbf{O}), P(-|\mathbf{S}), P(-|-)$[3].

Although this generative approach outperforms several models in correlation with coherence ratings assigned by human judges, it suffers in comparison with later systems. Barzilay and Lapata (2005) uses the same grid representation, but treats the transition probabilities $P(r_{i,j}|\vec{r}_{i,j})$ for each document as features for input to an SVM classifier. Soricut and Marcu (2006)'s implementation of the entity-based model also uses discriminative training.

The generative model's main weakness in comparison to these conditional models is its assumption of independence between entities. In real documents, each sentence tends to contain only a few nouns, and even fewer of them can fill roles like subject and object. In other words, nouns compete with each other for the available syntactic positions in sentences; once one noun is chosen as the subject, the probability that any other will also become a subject (of a different subclause of the same sentence) is drastically lowered. Since the generative entity grid does not take this into account, it learns that in general, the probability of any given entity appearing in a specific sentence is low. Thus it generates blank sentences (those without any nouns at all) with overwhelmingly high probability.

It may not be obvious that this misallocation of probability mass also reduces the effectiveness of the generative entity grid in ordering fixed sets of sentences. However, consider the case where an entity has a history $\vec{r}^h$, and then does not appear in the next sentence. The model treats this as evidence that entities generally do not occur immediately after $\vec{r}^h$– but it may also happen that the entity was outcompeted by some other word with even more significance.

## 3 Relaxed Entity Grid

In this section, we relax the troublesome assumption of independence between entities, thus moving the probability distribution over documents away from blank sentences. We begin at the same point as above: sequential generation of sentences: $P(D) = \prod_i P(S_i|S_{0..(i-1)})$. We similarly separate the words into entities and non-entities, treat the non-entities as independent of the history $\vec{S}$ and omit them. We also distinguish two types of entities. Let the *known set* $K_i = e_j : e_j \in \vec{S}_{(i-1)}$, the set of all entities which have appeared before sentence $i$. Of the entities appearing in $S_i$, those in $K_i$ are *known entities*, and those which are not are *new entities*. Since each entity in the document is new precisely once, we treat these as independent and omit them from our calculations as we did the non-entities. We return to both groups of omitted words in section 4 below when discussing our topic-based models.

To model a sentence, then, we generate the set of known entities it contains along with their syntactic roles, given the history and the known set $K_i$. We truncate the history, as above, with horizon $h$; note that this does not make the model Markovian, since the known set has no horizon. Finally, we consider only the portion of the history which relates to known nouns (since all non-known nouns have the same history - -). In all the equations below, we restrict $E_i$ to known entities which actually appear in sentence $i$, and $R_i$ to roles filled by known entities. The probability of a sentence is now:

$$P(S_i|\vec{S}^h_{(i-1)}) = P(E_i, R_i|\vec{R}^h_{(i-1)}). \quad (4)$$

We make one further simplification before beginning to approximate: we first generate the set of syntactic slots $R_i$ which we intend to fill with known entities, and then decide which entities from the known

set to select. Again, we assume independence from the history, so that the contribution of $P(R_i)$ for any ordering of a fixed set of sentences is constant and we omit it:

$$P(E_i, R_i|\vec{R}^h_{(i-1),j}) = P(E_i|R_i, \vec{R}^h_{(i-1),j}). \quad (5)$$

Estimating $P(E_i|R_i, \vec{R}^h_{(i-1),j})$ proves to be difficult, since the contexts are very sparse. To continue, we make a series of approximations. First let each role be filled individually (where $r \leftarrow e$ is the boolean indicator function "noun $e$ fills role $r$"):

$$P(E_i|R_i, \vec{R}^h_{(i-1),j}) \approx \prod_{r \in R_i} P(r \leftarrow e_j|r, \vec{R}^h_{(i-1),j}). \quad (6)$$

Notice that this process can select the same noun $e_j$ to fill multiple roles $r$, while the entity grid cannot represent such an occurrence. The resulting distribution is therefore slightly deficient.

Unfortunately, we are still faced with the sparse context $\vec{R}^h_{(i-1),j}$, the set of histories of all currently known nouns. It is much easier to estimate $P(r \leftarrow e_j|r, \vec{r}^h_{(i-1),j})$, where we condition only on the history of the particular noun which is chosen to fill slot $r$. However, in this case we do not have a proper probability distribution: i.e. the probabilities do not sum to 1. To overcome this difficulty we simply normalize by force[3]:

$$P(r \leftarrow e_j|r, \vec{R}^h_{(i-1),j}) \approx \frac{P(r \leftarrow e_j|r, \vec{r}^h_{(i-1),j})}{\sum_{j \in K_i} P(r \leftarrow e_j|r, \vec{r}^h_{(i-1),j})} \quad (7)$$

The individual probabilities $P(r \leftarrow e_j|r, \vec{r}^h_{(i-1),j})$ are calculated by counting situations in the training documents in which a known noun has history $\vec{r}^h_{(i-1),j}$ and fills slot $r$ in the next sentence, versus situations where the slot $r$ exists but is filled by some other noun. Some rare contexts are still sparse, and so we smooth by adding a pseudocount of 1 for all events. Our model is expressed by equations (1),(4),(5),(6) and (7). In figure 2, the probability of $S_3$ with horizon 1 is now calculated as follows: the known set contains PLAN, AIRPLANE, CONDITION, FLIGHT, PILOT, OWNER and OCCUPANT. There is one syntactic role filled by a known noun, **S**. The probability is then calculated as $P(+|\mathbf{S}, \mathbf{X})$ (the probability of selecting a noun with history **X** to fill the role of **S**) normalized by $P(+|\mathbf{S}, \mathbf{O}) + P(+|\mathbf{S}, \mathbf{S}) + P(+|\mathbf{S}, \mathbf{X}) + 4 \times P(+|\mathbf{S}, -)$.

Like Lapata and Barzilay (2005), our relaxed model assigns low probability to sentences where nouns with important-seeming histories do not appear. However, in our model, the penalty is less severe if there are many competitor nouns. On the other hand, if the sentence contains many slots, giving the noun more opportunity to fill one of them, the penalty is proportionally greater if it does not appear.

## 4 Topic-Based Model

The model we describe above is a purely local one, and moreover it relies on a particular set of local features which capture the way adjacent sentences tend to share lexical choices. Its lack of any global structure makes it impossible for the model to recover at a paragraph boundary, or to accurately guess which sentence should begin a document. Its lack of lexicalization, meanwhile, renders it incapable of learning dependences between pairs of words: for instance, that a sentence discussing a crash is often followed by a casualty report.

We remedy both these problems by extending our model of document generation. Like Barzilay and Lee (2004), we learn an HMM in which each sentence has a hidden topic $q_i$, which is chosen conditioned on the previous state $q_{i-1}$. The emission model of each state is an instance of the relaxed entity grid model as described above, but in addition to conditioning on the role and history, we condition also on the state and on the particular set of lexical items $lex(K_i)$ which may be selected to fill the role: $P(r \leftarrow e_j|r, \vec{R}^h_{(i-1),j}, q_i, lex(K_i))$. This distribution is approximated as above by the normalized value of $P(r \leftarrow e_j|r, \vec{r}^h_{(i-1),j}, q_i, lex(e_j))$. However, due to our use of lexical information, even this may be too sparse for accurate estimation, so we back off by interpolating with the pre-

[3] Unfortunately this estimator is not consistent (that is, given infinite training data produced by the model, the estimated parameters do not converge to the true parameters). We are investigating maximum entropy estimation as a solution to this problem.

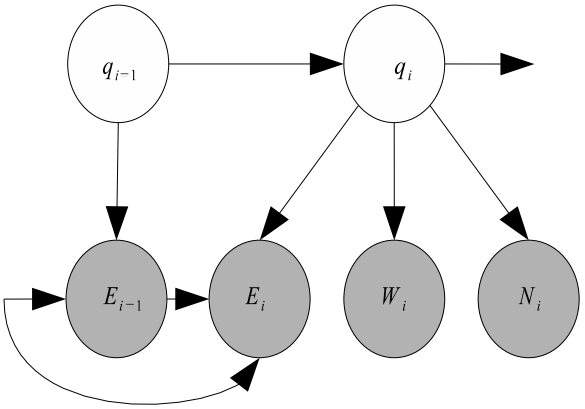


Figure 3: A single time-slice of our HMM.
$W_i \sim PY(\cdot|q_i; \theta_{LM}, discount_{LM})$
$N_i \sim PY(\cdot|q_i; \theta_{NN}, discount_{NN})$
$E_i \sim EGrid(\cdot|R, \vec{R}^2_{i-1}, q_i, lex(K_i); \theta_{EG})$
$q_i \sim DP(\cdot|q_{i-1})$
In the equations above, only the manually set interpolation hyperparameters are indicated.

vious model. In each context, we introduce $\theta_{EG}$ pseudo-observations, split fractionally according to the backoff distribution: if we abbreviate the context in the relaxed entity grid as $C$ and the event as $e$, this smoothing corresponds to:

$$P(e|C, q_i, e_j) = \frac{\#(e, C, q_i, e_j) + \theta_{EG} P(e|C)}{\#(e, C, q_i, e_j) + \theta_{EG}}.$$

This is equivalent to defining the topic-based entity grid as a Dirichlet process with parameter $\theta_{EG}$ sampling from the relaxed entity grid.

In addition, we are now in a position to generate the non-entity words $W_i$ and new entities $N_i$ in an informative way, by conditioning on the sentence topic $q_i$. Since they are interrupted by the known entities, they do not form contiguous sequences of words, so we make a bag-of-words assumption. To model these sets of words, we use unigram versions of the hierarchical Pitman-Yor processes of (Teh, 2006), which implement a Bayesian version of Kneser-Ney smoothing.

To represent the HMM itself, we adapt the nonparametric HMM of (Beal et al., 2001). This is a Bayesian alternative to the conventional HMM model learned using EM, chosen mostly for convenience. Our variant of it, unlike (Beal et al., 2001), has no parameter $\gamma$ to control self-transitions; our emission model is complex enough to make it unnecessary.

The actual number of states found by the model depends mostly on the backoff constants, the $\theta$s (and, for Pitman-Yor processes, $discount$s) chosen for the emission models (the entity grid, non-entity word model and new noun model), and is relatively insensitive to particular choices of prior for the other hyperparameters. As the backoff constants decrease, the emission models become more dependent on the state variable $q$, which leads to more states (and eventually to memorization of the training data). If instead the backoff rate increases, the emission models all become close to the general distribution and the model prefers relatively few states. We train with interpolations which generally result in around 40 states.

Once the interpolation constants are set, the model can be trained by Gibbs sampling. We also do inference over the remaining hyperparameters of the model by Metropolis sampling from uninformative priors. Convergence is generally very rapid; we obtain good results after about 10 iterations. Unlike Barzilay and Lee (2004), we do not initialize with an informative starting distribution.

When finding the probability of a test document, we do not do inference over the full Bayesian model, because the number of states, and the probability of different transitions, can change with every new observation, making dynamic programming impossible. Beal et al. (2001) proposes an inference algorithm based on particle filters, but we feel that in this case, the effects are relatively minor, so we approximate by treating the model as a standard HMM, using a fixed transition function based only on the training data. This allows us to use the conventional Viterbi algorithm. The backoff rates we choose at training time are typically too small for optimal inference in the ordering task. Before doing tests, we set them to higher values (determined to optimize ordering performance on held-out data) so that our emission distributions are properly smoothed.

## 5 Experiments

Our experiments use the popular AIRPLANE corpus, a collection of documents describing airplane crashes taken from the database of the National

Transportation Safety Board, used in (Barzilay and Lee, 2004; Barzilay and Lapata, 2005; Soricut and Marcu, 2006). We use the standard division of the corpus into 100 training and 100 test documents; for development purposes we did 10-fold cross-validation on the training data. The AIRPLANE documents have some advantages for coherence research: they are short (11.5 sentences on average) and quite formulaic, which makes it easy to find lexical and structural patterns. On the other hand, they do have some oddities. 46 of the training documents begin with a standard preamble: "This is preliminary information, subject to change, and may contain errors. Any errors in this report will be corrected when the final report has been completed," which essentially gives coherence models the first two sentences for free. Others, however, begin abruptly with no introductory material whatsoever, and sometimes without even providing references for their definite noun phrases; one document begins: "At V1, the DC-10-30's number 1 engine, a General Electric CF6-50C2, experienced a casing breach when the 2nd-stage low pressure turbine (LPT) anti-rotation nozzle locks failed." Even humans might have trouble identifying this sentence as the beginning of a document.

### 5.1 Sentence Ordering

In the sentence ordering task, (Lapata, 2003; Barzilay and Lee, 2004; Barzilay and Lapata, 2005; Soricut and Marcu, 2006), we view a document as an unordered bag of sentences and try to find the ordering of the sentences which maximizes coherence according to our model. This type of ordering process has applications in natural language generation and multi-document summarization. Unfortunately, finding the optimal ordering according to a probabilistic model with local features is NP-complete and non-approximable (Althaus et al., 2004). Moreover, since our model is not Markovian, the relaxation used as a heuristic for $A^*$ search by Soricut and Marcu (2006) is ineffective. We therefore use simulated annealing to find a high-probability ordering, starting from a random permutation of the sentences. Our search system has few Estimated Search Errors as defined by Soricut and Marcu (2006); it rarely proposes an ordering which has lower probability than the original ordering[4].

| | $\tau$ | Discr. (%) |
|---|---|---|
| (Barzilay and Lapata, 2005) | - | 90 |
| (Barzilay and Lee, 2004) | .44 | 74[5] |
| (Soricut and Marcu, 2006) | **.50** | -[6] |
| Topic-based (relaxed) | **.50** | **94** |

Table 1: Results for AIRPLANE test data.

To evaluate the quality of the orderings we predict as optimal, we use *Kendall's* $\tau$, a measurement of the number of pairwise swaps needed to transform our proposed ordering into the original document, normalized to lie between $-1$ (reverse order) and 1 (original order). Lapata (2006) shows that it corresponds well with human judgements of coherence and reading times. A slight problem with $\tau$ is that it does not always distinguish between proposed orderings of a document which disrupt local relationships at random, and orderings in which paragraph-like units move as a whole. In longer documents, it may be worth taking this problem into account when selecting a metric; however, the documents in the AIRPLANE corpus are mostly short and have little paragraph structure, so $\tau$ is an effective metric.

### 5.2 Discrimination

Our second task is the discriminative test used by (Barzilay and Lapata, 2005). In this task we generate random permutations of a test document, and measure how often the probability of a permutation is higher than that of the original document. This task bears some resemblance to the task of discriminating coherent from incoherent essays in (Miltsakaki and Kukich, 2004), and is also equivalent in the limit to the ranking metric of (Barzilay and Lee, 2004), which we cannot calculate because our model does not produce $k$-best output. As opposed to the ordering task, which tries to measure how *close* the model's preferred orderings are to the original, this measurement assesses how *many* orderings the model prefers. We use 20 random permutations per document, for 2000 total tests.

| | $\tau$ | Discr. (%) |
|---|---|---|
| Naive Entity Grid | .17 | 81 |
| Relaxed Entity Grid | .02 | 87 |
| Topic-based (naive) | .39 | 85 |
| Topic-based (relaxed) | **.54** | **96** |

Table 2: Results for 10-fold cross-validation on AIRPLANE training data.

## 6 Results

Since the ordering task requires a model to propose the complete structure for a set of sentences, it is very dependent on global features. To perform adequately, a model must be able to locate the beginning and end of the document, and place intermediate sentences relative to these two points. Without any way of doing this, our relaxed entity grid model has $\tau$ of approximately 0, meaning its optimal orderings are essentially uncorrelated with the correct orderings[7]. The HMM content model of (Barzilay and Lee, 2004), which does have global structure, performs much better on ordering, at $\tau$ of .44. However, local features can help substantially for this task, since models which use them are better at placing related sentences next to one another. Using both sets of features, our topic-based model achieves state of the art performance ($\tau = .5$) on the ordering task, comparable with the mixture model of (Soricut and Marcu, 2006).

The need for good local coherence features is especially clear from the results on the discrimination task (table 1). Permuting a document may leave obvious "signposts" like the introduction and conclusion in place, but it almost always splits up many pairs of neighboring sentences, reducing local coherence. (Barzilay and Lee, 2004), which lacks local features, does quite poorly on this task (74%), while our model performs extremely well (94%).

It is also clear from the results that our relaxed entity grid model (87%) improves substantially on the generative naive entity grid (81%). When used on its own, it performs much better on the discrimination task, which is the one for which it was designed. (The naive entity grid has a higher $\tau$ score, .17, essentially by accident. It slightly prefers to generate infrequent nouns from the start context rather than the context - -, which happens to produce the correct placement for the "preliminary information" preamble.) When used as the emission model for known entities in our topic-based system, the relaxed entity grid shows its improved performance even more strongly (table 2); its results are about 10% higher than the naive version under both metrics.

Our combined model uses only entity-grid features and unigram language models,a strict subset of the feature set of (Soricut and Marcu, 2006). Their mixture includes an entity grid model and a version of the HMM of (Barzilay and Lee, 2004), which uses n-gram language modeling. It also uses a model of lexical generation based on the IBM-1 model for machine translation, which produces all words in the document conditioned on words from previous sentences. In contrast, we generate only entities conditioned on words from previous sentences; other words are conditionally independent given the topic variable. It seems likely therefore that using our model as a component of a mixture might improve on the state of the art result.

## 7 Future Work

Ordering in the AIRPLANE corpus and similar constrained sets of short documents is by no means a solved problem, but the results so far show a good deal of promise. Unfortunately, in longer and less formulaic corpora, the models, inference algorithms and even evaluation metrics used thus far may prove extremely difficult to scale up. Domains with more natural writing styles will make lexical prediction a much more difficult problem. On the other hand, the wider variety of grammatical constructions used may motivate more complex syntactic features, for instance as proposed by (Siddharthan et al., 2004) in sentence clustering.

Finding optimal orderings is a difficult task even for short documents, and will become exponentially more challenging in longer ones. For multi-paragraph documents, it is probably impractical to use full-scale coherence models to find optimal or-

[4] 0 times on test data, 3 times in cross-validation.

[5] Calculated on our test permutations using the code at *http://people.csail.mit.edu/regina/code.html*.

[6] Soricut and Marcu (2006) do not report results on this task, except to say that their implementation of the entity grid performs comparably to (Barzilay and Lapata, 2005).

[7] Barzilay and Lapata (2005) do not report $\tau$ scores.

derings directly. A better approach may be a coarse-to-fine or hierarchical system which cuts up longer documents into more manageable chunks that can be ordered as a unit.

Multi-paragraph documents also pose a problem for the $\tau$ metric itself. In documents with clear thematic divisions between their different sections, a good ordering metric should treat transposed paragraphs differently than transposed sentences.

## 8 Acknowledgements

We are extremely grateful to Regina Barzilay, for her code, data and extensive support, Mirella Lapata for code and advice, and the BLLIP group, especially Tom Griffiths, Sharon Goldwater and Mark Johnson, for comments and criticism. We were supported by DARPA GALE contract HR0011-06-2-0001 and the Karen T. Romer Foundation. Finally we thank three anonymous reviewers for their comments.

## References

Ernst Althaus, Nikiforos Karamanis, and Alexander Koller. 2004. Computing locally coherent discourses. In *Proceedings of the 42nd ACL*, Barcelona.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: an entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

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

Matthew J. Beal, Zoubin Ghahramani, and Carl Edward Rasmussen. 2001. The infinite Hidden Markov Model. In *NIPS*, pages 577–584.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine *n*-best parsing and MaxEnt discriminative reranking. In *Proc. of the 2005 Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 173–180.

Peter Foltz, Walter Kintsch, and Thomas Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2&3):285–307.

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

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *HLT-NAACL*, pages 185–192.

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

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, pages 1085–1090.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the annual meeting of ACL, 2003*.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):1–14.

E. Miltsakaki and K. Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Nat. Lang. Eng.*, 10(1):25–55.

Advaith Siddharthan, Ani Nenkova, and Kathleen McKeown. 2004. Syntactic simplification for improving content selection in multi-document summarization. In *COLING04*, pages 896–902.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the Association for Computational Linguistics Conference (ACL-2006)*.

Y.W. Teh. 2006. A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, National University of Singapore.