Disentangling Chat with Local Coherence Models

Micha Elsner
School of Informatics
University of Edinburgh
melsner0@gmail.com

Eugene Charniak
Department of Computer Science
Brown University, Providence, RI 02912
ec@cs.brown.edu

Abstract

We evaluate several popular models of local discourse coherence for domain and task generality by applying them to chat disentanglement. Using experiments on synthetic multiparty conversations, we show that most models transfer well from text to dialogue. Coherence models improve results overall when good parses and topic models are available, and on a constrained task for real chat data.

1 Introduction

One property of a well-written document is *coherence*, the way each sentence fits into its context– sentences should be interpretable in light of what has come before, and in turn make it possible to interpret what comes after. Models of coherence have primarily been used for text-based generation tasks: ordering units of text for multidocument summarization or inserting new text into an existing article. In general, the corpora used consist of informative writing, and the tasks used for evaluation consider different ways of reordering the same set of textual units. But the theoretical concept of coherence goes beyond both this domain and this task setting– and so should coherence models.

This paper evaluates a variety of local coherence models on the task of chat disentanglement or "threading": separating a transcript of a multiparty interaction into independent conversations[1]. Such simultaneous conversations occur in internet chat rooms, and on shared voice channels such as push-to-talk radio. In these situations, a single, correctly disentangled, conversational thread will be coherent, since the speakers involved understand the normal rules of discourse, but the transcript as a whole will not be. Thus, a good model of coherence should be able to disentangle sentences as well as order them.

There are several differences between disentanglement and the newswire sentence-ordering tasks typically used to evaluate coherence models. Internet chat comes from a different domain, one where topics vary widely and no reliable syntactic annotations are available. The disentanglement task measures different capabilities of a model, since it compares documents that are not permuted versions of one another. Finally, full disentanglement requires a large-scale search, which is computationally difficult. We move toward disentanglement in stages, carrying out a series of experiments to measure the contribution of each of these factors.

As an intermediary between newswire and internet chat, we adopt the SWITCHBOARD (SWBD) corpus. SWBD contains recorded telephone conversations with known topics and hand-annotated parse trees; this allows us to control for the performance of our parser and other informational resources. To compare the two algorithmic settings, we use SWBD for ordering experiments, and also artificially entangle pairs of telephone dialogues to create synthetic transcripts which we can disentangle. Finally, we present results on actual internet chat corpora.

On synthetic SWBD transcripts, local coherence models improve performance considerably over our baseline model, Elsner and Charniak (2008b). On

[1] A public implementation is available via https://bitbucket.org/melsner/browncoherence.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 1179–1189,
Portland, Oregon, June 19-24, 2011.

internet chat, we continue to do better on a constrained disentanglement task, though so far, we are unable to apply these improvements to the full task. We suspect that, with better low-level annotation tools for the chat domain and a good way of integrating prior information, our improvements on SWBD could transfer fully to IRC chat.

2 Related work

There is extensive previous work on coherence models for text ordering; we describe several specific models below, in section 2. This study focuses on models of local coherence, which relate text to its immediate context. There has also been work on global coherence, the structure of a document as a whole (Chen et al., 2009; Eisenstein and Barzilay, 2008; Barzilay and Lee, 2004), typically modeled in terms of sequential topics. We avoid using them here, because we do not believe topic sequences are predictable in conversation and because such models tend to be algorithmically cumbersome.

In addition to text ordering, local coherence models have also been used to score the fluency of texts written by humans or produced by machine (Pitler and Nenkova, 2008; Lapata, 2006; Miltsakaki and Kukich, 2004). Like disentanglement, these tasks provide an algorithmic setting that differs from ordering, and so can demonstrate previously unknown weaknesses in models. However, the target genre is still informative writing, so they reveal little about cross-domain flexibility.

The task of disentanglement or "threading" for internet chat was introduced by Shen et al. (2006). Elsner and Charniak (2008b) created the publicly available #LINUX corpus; the best published results on this corpus are those of Wang and Oard (2009). These two studies use overlapping unigrams to measure similarity between two sentences; Wang and Oard (2009) use a message expansion technique to incorporate context beyond a single sentence. Unigram overlaps are used to model coherence, but more sophisticated methods using syntax (Lapata and Barzilay, 2005) or lexical features (Lapata, 2003) often outperform them on ordering tasks. This study compares several of these methods with Elsner and Charniak (2008b), which we use as a baseline because there is a publicly available implementation[2].

Adams (2008) also created and released a disentanglement corpus. They use LDA (Blei et al., 2001) to discover latent topics in their corpus, then measuring similarity by looking for shared topics. These features fail to improve their performance, which is puzzling in light of the success of topic modeling for other coherence and segmentation problems (Eisenstein and Barzilay, 2008; Foltz et al., 1998). The results of this study suggest that topic models can help with disentanglement, but that it is difficult to find useful topics for IRC chat.

A few studies have attempted to disentangle conversational speech (Aoki et al., 2003; Aoki et al., 2006), mostly using temporal features. For the most part, however, this research has focused on auditory processing in the context of the *cocktail party* problem, the task of attending to a specific speaker in a noisy room (Haykin and Chen, 2005). Utterance content has some influence on what the listener perceives, but only for extremely salient cues such as the listener's name (Moray, 1959), so cocktail party research does not typically use lexical models.

3 Models

In this section, we briefly describe the models we intend to evaluate. Most of them are drawn from previous work; one, the topical entity grid, is a novel extension of the entity grid. For the experiments below, we train the models on SWBD, sometimes augmented with a larger set of automatically parsed conversations from the FISHER corpus. Since the two corpora are quite similar, FISHER is a useful source for extra data; McClosky et al. (2010) uses it for this purpose in parsing experiments. (We continue to use SWBD/FISHER even for experiments on IRC, because we do not have enough disentangled training data to learn lexical relationships.)

3.1 Entity grid

The entity grid (Lapata and Barzilay, 2005; Barzilay and Lapata, 2005) is an attempt to model some principles of Centering Theory (Grosz et al., 1995) in a statistical manner. It represents a document in terms of entities and their syntactic roles: subject (S), object (O), other (X) and not present (-). In each new

[2] cs.brown.edu/~melsner

utterance, the grid predicts the role in which each entity will appear, given its history of roles in the previous sentences, plus a *salience* feature counting the total number of times the entity occurs. For instance, for an entity which is the subject of sentence 1, the object of sentence 2, and occurs four times in total, the grid predicts its role in sentence 3 according to the conditional $P(\cdot|S, O, sal = 4)$.

As in previous work, we treat each noun in a document as denoting a single entity, rather than using a coreference technique to attempt to resolve them. In our development experiments, we noticed that coreferent nouns often occur farther apart in conversation than in newswire, since they are frequently referred to by pronouns and deictics in the interim. Therefore, we extend the history to six previous utterances. For robustness with this long history, we model the conditional probabilities using multilabel logistic regression rather than maximum likelihood. This requires the assumption of a linear model, but makes the estimator less vulnerable to overfitting due to sparsity, increasing performance by about 2% in development experiments.

3.2 Topical entity grid

This model is a variant of the generative entity grid, intended to take into account topical information. To create the topical entity grid, we learn a set of topic-to-word distributions for our corpus using LDA (Blei et al., 2001)[3] with 200 latent topics. This model embeds our vocabulary in a low-dimensional space: we represent each word w as the vector of topic probabilities $p(t_i|w)$. We experimented with several ways to measure relationships between words in this space, starting with the standard cosine. However, the cosine can depend on small variations in probability (for instance, if w has most of its mass in dimension 1, then it is sensitive to the exact weight of v for topic 1, even if this essentially never happens).

To control for this tendency, we instead use the magnitude of the dimension of greatest similarity:

$$sim(w, v) = max_i \; min(w_i, v_i)$$

To model coherence, we generalize the binary history features of the standard entity grid, which detect, for example, whether entity e is the subject of the previous sentence. In the topical entity grid, we instead compute a real-valued feature which sums up the similarity between entity e and the subject(s) of the previous sentence.

These features can detect a transition like: "The House voted yesterday. The Senate will consider the bill today.". If "House" and "Senate" have a high similarity, then the feature will have a high value, predicting that "Senate" is a good subject for the current sentence. As in the previous section, we learn the conditional probabilities with logistic regression; we train in parallel by splitting the data and averaging (Mann et al., 2009). The topics are trained on FISHER, and on NANC for news.

3.3 IBM-1

The IBM translation model was first considered for coherence by Soricut and Marcu (2006), although a less probabilistically elegant version was proposed earlier (Lapata, 2003). This model attempts to generate the content words of the next sentence by translating them from the words of the previous sentence, plus a null word; thus, it will learn alignments between pairs of words that tend to occur in adjacent sentences. We learn parameters on the FISHER corpus, and on NANC for news.

3.4 Pronouns

The use of a generative pronoun resolver for coherence modeling originates in Elsner and Charniak (2008a). That paper used a supervised model (Ge et al., 1998), but we adapt a newer, unsupervised model which they also make publicly available (Charniak and Elsner, 2009)[4]. They model each pronoun as generated by an antecedent somewhere in the previous two sentences. If a good antecedent is found, the probability of the pronoun's occurrence will be high; otherwise, the probability is low, signaling that the text is less coherent because the pronoun is hard to interpret correctly.

We use the model as distributed for news text. For conversation, we adapt it by running a few iterations of their EM training algorithm on the FISHER data.

[3] www.cs.princeton.edu/~blei/topicmodeling.html

[4] bllip.cs.brown.edu/resources.shtml\#software

3.5 Discourse-newness

Building on work from summarization (Nenkova and McKeown, 2003) and coreference resolution (Poesio et al., 2005), Elsner and Charniak (2008a) use a model which recognizes discourse-new versus old NPs as a coherence model. For instance, the model can learn that "President Barack Obama" is a more likely first reference than "Obama". Following their work, we score discourse-newness with a maximum-entropy classifier using syntactic features counting different types of NP modifiers, and we use NP head identity as a proxy for coreference.

3.6 Chat-specific features

Most disentanglement models use non-linguistic information alongside lexical features; in fact, timestamps and speaker identities are usually *better* cues than words are. We capture three essential non-linguistic features using simple generative models.

The first feature is the **time** gap between one utterance and the next within the same thread. Consistent short gaps are a sign of normal turn-taking behavior; long pauses do occur, but much more rarely (Aoki et al., 2003). We round all time gaps to the nearest second and model the distribution of time gaps using a histogram, choosing bucket sizes adaptively so that each bucket contains at least four datapoints.

The second feature is **speaker** identity; conversations usually involve a small subset of the total number of speakers, and a few core speakers make most of the utterances. We model the distribution of speakers in each conversation using a Chinese Restaurant Process (CRP) (Aldous, 1985) (tuning the dispersion α to maximize development performance). The CRP's "rich-get-richer" dynamics capture our intuitions, favoring conversations dominated by a few vociferous speakers.

Finally, we model name **mentioning**. Speakers in IRC chat often use their addressee's names to coordinate the chat (O'Neill and Martin, 2003), and this is a powerful source of information (Elsner and Charniak, 2008b). Our model classifies each utterance into either the start or continuation of a conversational turn, by checking if the previous utterance had the same speaker. Given this status, it computes probabilities for three outcomes: no name mention, a mention of someone who has previously spoken in the conversation, or a mention of someone else. (The third option is extremely rare; this accounts for most of the model's predictive power). We learn these probabilities from IRC training data.

3.7 Model combination

To combine these different models, we adopt the log-linear framework of Soricut and Marcu (2006). Here, each model P_i is assigned a weight λ_i, and the combined score $P(d)$ is proportional to:

$$\sum_i \lambda_i log(P_i(d))$$

The weights λ can be learned discriminatively, maximizing the probability of d relative to a task-specific contrast set. For ordering experiments, the contrast set is a single random permutation of d; we explain the training regime for disentanglement below, in subsection 4.1.

4 Comparing orderings of SWBD

To measure the differences in performance caused by moving from news to a conversational domain, we first compare our models on an ordering task, discrimination (Barzilay and Lapata, 2005; Karamanis et al., 2004). In this task, we take an original document and randomly permute its sentences, creating an artificial incoherent document. We then test to see if our model prefers the coherent original.

For SWBD, rather than compare permutations of the individual utterances, we permute conversational turns (sets of consecutive utterances by each speaker), since turns are natural discourse units in conversation. We take documents numbered 2000–3999 as training/development and the remainder as test, yielding 505 training and 153 test documents; we evaluate 20 permutations per document. As a comparison, we also show results for the same models on WSJ, using the train-test split from Elsner and Charniak (2008a); the test set is sections 14-24, totalling 1004 documents.

Purandare and Litman (2008) carry out similar experiments on distinguishing permuted SWBD documents, using lexical and WordNet features in a model similar to Lapata (2003). Their accuracy for this task (which they call "switch-hard") is roughly 68%.

	WSJ	SWBD
EGrid	76.4‡	86.0
Topical EGrid	71.8‡	70.9‡
IBM-1	77.2‡	84.9†
Pronouns	69.6‡	71.7‡
Disc-new	72.3‡	55.0‡
Combined	81.9	88.4
-EGrid	81.0	87.5
-Topical EGrid	**82.2**	**90.5**
-IBM-1	79.0‡	88.9
-Pronouns	81.3	88.5
-Disc-new	**82.2**	88.4

Table 1: Discrimination F scores on news and dialogue. ‡ indicates a significant difference from the combined model at p=.01 and † at p=.05.

In Table 1, we show the results for individual models, for the combined model, and ablation results for mixtures without each component. WSJ is more difficult than SWBD overall because, on average, news articles are shorter than SWBD conversations. Short documents are harder, because permuting disrupts them less. The best SWBD result is 91%; the best WSJ result is 82% (both for mixtures without the topical entity grid). The WSJ result is state-of-the-art for the dataset, improving slightly on Elsner and Charniak (2008a) at 81%. We test results for significance using the non-parametric Mann-Whitney U test.

Controlling for the fact that discrimination is easier on SWBD, most of the individual models perform similarly in both corpora. The strongest models in both cases are the entity grid and IBM-1 (at about 77% for news, 85% for dialogue). Pronouns and the topical entity grid are weaker. The major outlier is the discourse-new model, whose performance drops from 72% for news to only 55%, just above chance, for conversation.

The model combination results show that all the models are quite closely correlated, since leaving out any single model does not degrade the combination very much (only one of the ablations is significantly worse than the combination). The most critical in news is IBM-1 (decreasing performance by 3% when removed); in conversation, it is the entity grid (decreasing by about 1%). The topical entity grid actually has a (nonsignificant) *negative* impact on combined performance, implying that its predictive power in this setting comes mainly from information that other models also capture, but that it is noisier and less reliable. In each domain, the combined models outperform the best single model, showing the information provided by the weaker models is not completely redundant.

Overall, these results suggest that most previously proposed local coherence models are domain-general; they work on conversation as well as news. The exception is the discourse-newness model, which benefits most from the specific conventions of a written style. Full names with titles (like "President Barack Obama") are more common in news, while conversation tends to involve fewer completely unfamiliar entities and more cases of bridging reference, in which grounding information is given implicitly (Nissim, 2006). Due to its poor performance, we omit the discourse-newness model in our remaining experiments.

5 Disentangling SWBD

We now turn to the task of disentanglement, testing whether models that are good at ordering also do well in this new setting. We would like to hold the domain constant, but we do not have any disentanglement data recorded from naturally occurring speech, so we create synthetic instances by merging pairs of SWBD dialogues. Doing so creates an artificial transcript in which two pairs of people appear to be talking simultaneously over a shared channel.

The situation is somewhat contrived in that each pair of speakers converses only with each other, never breaking into the other pair's dialogue and rarely using devices like name mentioning to make it clear who they are addressing. Since this makes speaker identity a perfect cue for disentanglement, we do not use it in this section. The only chat-specific model we use is **time**.

Because we are not using speaker information, we remove all utterances which do not contain a noun before constructing synthetic transcripts– these are mostly backchannels like "Yeah". Such utterances cannot be correctly assigned by our coherence models, which deal with content; we suspect most of them could be dealt with by associating them with the nearest utterance from the same speaker.

Once the backchannels are stripped, we can create a synthetic transcript. For each dialogue, we first simulate timestamps by sampling the number of seconds between each utterance and the next from a discretized Gaussian: $\lfloor N(0, 2.5) \rfloor$. The interleaving of the conversations is dictated by the timestamps. We truncate the longer conversation at the length of the shorter; this ensures a baseline score of 50% for the degenerate model that assigns all utterances to the same conversation.

We create synthetic instances of two types– those where the two entangled conversations had different topical prompts and those where they were the same. (Each dialogue in SWBD focuses on a preselected topic, such as *fishing* or *movies*.) We entangle dialogues from our ordering development set to use for mixture training and validation; for testing, we use 100 instances of each type, constructed from dialogues in our test set.

When disentangling, we treat each thread as independent of the others. In other words, the probability of the entire transcript is the product of the probabilities of the component threads. Our objective is to find the set of threads maximizing this. As a comparison, we use the model of Elsner and Charniak (2008b) as a baseline. To make their implementation comparable to ours, in this section we constrain it to find only two threads.

5.1 Disentangling a single utterance

Our first disentanglement task is to correctly assign a single utterance, given the true structure of the rest of the transcript. For each utterance, we compare two versions of the transcript, the original, and a version where it is swapped into the other thread. Our accuracy measures how often our models prefer the original. Unlike full-scale disentanglement, this task does not require a computationally demanding search, so it is possible to run experiments quickly. We also use it to train our mixture models for disentanglement, by construct a training example for each utterance i in our training transcripts. Since the Elsner and Charniak (2008b) model maximizes a correlation clustering objective which sums up independent edge weights, we can also use it to disentangle a single sentence efficiently.

Our results are shown in Table 2. Again, results for individual models are above the line, then our combined model, and finally ablation results for mixtures omitting a single model. The results show that, for a pair of dialogues that differ in topic, our best model can assign a single sentence with 87% accuracy. For the same topic, the accuracy is 80%. In each case, these results improve on (Elsner and Charniak, 2008b), which scores 78% and 74%.

	Different	Same	Avg.
EGrid	80.2	72.9	76.6
Topical EGrid	81.7	73.3	77.5
IBM-1	70.4	66.7	68.5
Pronouns	53.1	50.1	51.6
Time	58.5	57.4	57.9
Combined	**86.8**	**79.6**	**83.2**
-EGrid	86.0	79.1	82.6
-Topical EGrid	85.2	78.7	81.9
-IBM-1	86.2	78.7	82.4
-Pronouns	**86.8**	79.4	83.1
-Time	84.5	76.7	80.6
E+C '08	78.2	73.5	75.8

Table 2: Average accuracy for disentanglement of a single utterance on 200 synthetic multiparty conversations from SWBD test.

Changing to this new task has a substantial impact on performance. The topical model, which performed poorly for ordering, is actually stronger than the entity grid in this setting. IBM-1 underperforms either grid model (69% to 77%); on ordering, it was nearly as good (85% to 86%).

Despite their ordering performance of 72%, pronouns are essentially useless for this task, at 52%. This decline is due partly to domain, and partly to task setting. Although SWBD contains more pronominals than WSJ, many of them are first and second-person pronouns or deictics, which our model does not attempt to resolve. Since the disentanglement task involves moving only a single sentence, if moving this sentence does not sever a resolvable pronoun from its antecedent, the model will be unable to make a good decision.

As before, the ablation results show that all the models are quite correlated, since removing any single model from the mixture causes only a small decrease in performance. The largest drop (83% to 81%) is caused by removing time; though time is a weak model on its own, it is completely orthogo-

nal to the other models, since unlike them, it does not depend on the words in the sentences.

Comparing results between "different topic" and "same topic" instances shows that "same topic" is harder– by about 7% for the combined model. The IBM model has a relatively small gap of 3.7%, and in the ablation results, removing it causes a larger drop in performance for "same" than "different"; this suggests it is somewhat more robust to similarity in topic than entity grids.

Disentanglement accuracy is hard to predict given ordering performance; the two tasks plainly make different demands on models. One difference is that the models which use longer histories (the two entity grids) remain strong, while the models considering only one or two previous sentences (IBM and pronouns) do not do as well. Since the changes being considered here affect only a single sentence, while permutation affects the entire transcript, more history may help by making the model more sensitive to small changes.

5.2 Disentangling an entire transcript

We now turn to the task of disentangling an entire transcript at once. This is a practical task, motivated by applications such as search and information retrieval. However, it is more difficult than assigning only a single utterance, because decisions are interrelated– an error on one utterance may cause a cascade of poor decisions further down. It is also computationally harder.

We use tabu search (Glover and Laguna, 1997) to find a good solution. The search repeatedly finds and moves the utterance which would most improve the model score if swapped from one thread to the other. Unlike greedy search, tabu search is constrained not to repeat a solution that it has recently visited; this forces it to keep exploring when it reaches a local maximum. We run 500 iterations of tabu search (usually finding the first local maximum after about 100) and return the best solution found.

We measure performance with one-to-one overlap, which maps the two clusters to the two gold dialogues, then measures percent correct[5]. Our results (Table 3) show that, for transcripts with different topics, our disentanglement has 68% overlap with truth, extracting about two thirds of the structure correctly; this is substantially better than Elsner and Charniak (2008b), which scores 59%. Where the entangled conversations have the same topic, performance is lower, about 60%, but still better than the comparison model with 57%. Since correlations with the previous section are fairly reliable, and the disentanglement procedure is computationally intensive, we omit ablation experiments.

	Different	Same	Avg.
EGrid	60.3	57.1	58.7
Topical EGrid	62.3	56.8	59.6
IBM-1	56.5	55.2	55.9
Pronouns	54.5	54.4	54.4
Time	55.4	53.8	54.6
Combined	**67.9**	**59.8**	**63.9**
E+C '08	59.1	57.4	58.3

Table 3: One-to-one overlap between disentanglement results and truth on 200 synthetic multiparty conversations from SWBD test.

As we expect, full disentanglement is more difficult than single-sentence disentanglement (combined scores drop by about 20%), but the single-sentence task is a good predictor of relative performance. Entity grid models do best, the IBM model remains useful, but less so than for discrimination, and pronouns are very weak. The IBM model performs similarly under both metrics (56% and 57%), while other models perform worse on loc_3. This supports our suggestion that IBM's decline in performance from ordering is indeed due to its using a single sentence history; it is still capable of getting local structures right, but misses global ones.

6 IRC data

In this section, we move from synthetic data to real multiparty discourse recorded from internet chat rooms. We use two datasets: the #LINUX corpus (Elsner and Charniak, 2008b), and three larger corpora, #IPHONE, #PHYSICS and #PYTHON (Adams, 2008). We use the 1000-line "development" section of #LINUX for tuning our mixture models and the 800-line "test" section for development experiments. We reserve the Adams (2008) corpora for testing; together, they consist of 19581 lines of chat, with each section containing 500 to 1000 lines.

[5]The other popular metrics, F and loc_3, are correlated.

Chat-specific	74.0
+EGrid	**79.3**
+Topical EGrid	76.8
+IBM-1	76.3
+Pronouns	73.9
+EGrid/Topic/IBM-1	78.3
E+C '08b	76.4

Table 4: Accuracy for single utterance disentanglement, averaged over annotations of 800 lines of #LINUX data.

In order to use syntactic models like the entity grid, we parse the transcripts using (McClosky et al., 2006). Performance is bad, although the parser does identify most of the NPs; poor results are typical for a standard parser on chat (Foster, 2010). We postprocess the parse trees to retag "lol", "haha" and "yes" as UH (rather than NN, NNP and JJ).

In this section, we use all three of our chat-specific models (sec. 2.0.6; **time**, **speaker** and **mention**) as a baseline. This baseline is relatively strong, so we evaluate our other models in combination with it.

6.1 Disentangling a single sentence

As before, we show results on correctly disentangling a single sentence, given the correct structure of the rest of the transcript. We average performance on each transcript over the different annotations, then average the transcripts, weighing them by length to give each utterance equal weight.

Table 4 gives results on our development corpus, #LINUX. Our best result, for the chat-specific features plus entity grid, is 79%, improving on the comparison model, Elsner and Charniak (2008b), which gets 76%. (Although the table only presents an average over all annotations of the dataset, this model is also more accurate for each individual annotator than the comparison model.) We then ran the same model, chat-specific features plus entity grid, on the test corpora from Adams (2008). These results (Table 5) are also better than Elsner and Charniak (2008b), at an average of 93% over 89%.

As pointed out in Elsner and Charniak (2008b), the chat-specific features are quite powerful in this domain, and it is hard to improve over them. Elsner and Charniak (2008b), which has simple lexical features, mostly based on unigram overlap, increases

	#IPHONE	#PHYSICS	#PYTHON
+EGrid	**92.3**	**96.6**	**91.1**
E+C '08b	89.0	90.2	88.4

Table 5: Average accuracy for disentanglement of a single utterance for 19581 total lines from Adams (2008).

performance over baseline by 2%. Both IBM and the topical entity grid achieve similar gains. The entity grid does better, increasing performance to 79%. Pronouns, as before for SWBD, are useless.

We believe that the entity grid's good performance here is due mostly to two factors: its use of a long history, and its lack of lexicalization. The grid looks at the previous six sentences, which differentiates it from the IBM model and from Elsner and Charniak (2008b), which treats each pair of sentences independently. Using this long history helps to distinguish important nouns from unimportant ones better than frequency alone. We suspect that our lexicalized models, IBM and the topical entity grid, are hampered by poor parameter settings, since their parameters were learned on FISHER rather than IRC chat. In particular, we believe this explains why the topical entity grid, which slightly outperformed the entity grid on SWBD, is much worse here.

6.2 Full disentanglement

Running our tabu search algorithm on the full disentanglement task yields disappointing results. Accuracies on the #LINUX dataset are not only worse than previous work, but also worse than simple baselines like creating one thread for each speaker. The model finds far too many threads– it detects over 300, when the true number is about 81 (averaging over annotations). This appears to be related to biases in our chat-specific models as well as in the entity grid; the time model (which generates gaps between adjacent sentences) and the speaker model (which uses a CRP) both assign probability 1 to single-utterance conversations. The entity grid also has a bias toward short conversations, because unseen entities are empirically more likely to occur toward the beginning of a conversation than in the middle.

A major weakness in our model is that we aim only to maximize coherence of the individual conversations, with no prior on the likely length or number of conversations that will appear in the tran-

script. This allows the model to create far too many conversations. Integrating a prior into our framework is not straightforward because we currently train our mixture to maximize single-utterance disentanglement performance, and the prior is not useful for this task.

We experimented with fixing parts of the transcript to the solution obtained by Elsner and Charniak (2008b), then using tabu search to fill in the gaps. This constrains the number of conversations and their approximate positions. With this structure in place, we were able to obtain scores comparable to Elsner and Charniak (2008b), but not improvements. It appears that our performance increase on single-sentence disentanglement does not transfer to this task because of cascading errors and the necessity of using external constraints.

7 Conclusions

We demonstrate that several popular models of local coherence transfer well to the conversational domain, suggesting that they do indeed capture coherence in general rather than specific conventions of newswire text. However, their performance across tasks is not as stable; in particular, models which use less history information are worse for disentanglement.

Our results study suggest that while sophisticated coherence models can potentially contribute to disentanglement, they would benefit greatly from improved low-level resources for internet chat. Better parsing, or at least NP chunking, would help for models like the entity grid which rely on syntactic role information. Larger training sets, or some kind of transfer learning, could improve the learning of topics and other lexical parameters. In particular, our results on SWBD data confirm the conjecture of (Adams, 2008) that LDA topic modeling is in principle a useful tool for disentanglement– we believe a topic-based model could also work on IRC chat, but would require a better set of extracted topics. With better parameters for these models and the integration of a prior, we believe that our good performance on SWBD and single-utterance disentanglement for IRC can be extended to full-scale disentanglement of IRC.

Acknowledgements

We are extremely grateful to Regina Barzilay, Mark Johnson, Rebecca Mason, Ben Swanson and Neal Fox for their comments, to Craig Martell for the NPS chat datasets and to three anonymous reviewers. This work was funded by a Google Fellowship for Natural Language Processing.

References

Paige H. Adams. 2008. *Conversation Thread Extraction and Topic Detection in Text-based Chat.* Ph.D. thesis, Naval Postgraduate School.

David Aldous. 1985. Exchangeability and related topics. In *Ecole d'Ete de Probabilities de Saint-Flour XIII 1983*, pages 1–198. Springer.

Paul M. Aoki, Matthew Romaine, Margaret H. Szymanski, James D. Thornton, Daniel Wilson, and Allison Woodruff. 2003. The mad hatter's cocktail party: a social mobile audio space supporting multiple simultaneous conversations. In *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 425–432, New York, NY, USA. ACM Press.

Paul M. Aoki, Margaret H. Szymanski, Luke D. Plurkowski, James D. Thornton, Allison Woodruff, and Weilie Yi. 2006. Where's the "party" in "multiparty"?: analyzing the structure of small-group sociable talk. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 393–402, New York, NY, USA. ACM Press.

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.

David Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:2003.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of EACL*, Athens, Greece.

Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. 2009. Global models of document structure using latent permutations. In *Proceedings of Human Language Technologies: The 2009 Annual*

Conference of the North American Chapter of the Association for Computational Linguistics, pages 371–379, Boulder, Colorado, June. Association for Computational Linguistics.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *EMNLP*, pages 334–343.

Micha Elsner and Eugene Charniak. 2008a. Coreference-inspired coherence modeling. In *Proceedings of ACL-08: HLT, Short Papers*, pages 41–44, Columbus, Ohio, June. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2008b. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Columbus, Ohio, June. Association for Computational Linguistics.

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

Jennifer Foster. 2010. "cba to check the spelling": Investigating parser performance on discussion forum posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384, Los Angeles, California, June. Association for Computational Linguistics.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–171, Orlando, Florida. Harcourt Brace.

Fred Glover and Manuel Laguna. 1997. *Tabu Search.* University of Colorado at Boulder.

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.

Simon Haykin and Zhe Chen. 2005. The Cocktail Party Problem. *Neural Computation*, 17(9):1875–1902.

Nikiforos Karamanis, Massimo Poesio, Chris Mellish, and Jon Oberlander. 2004. Evaluating centering-based metrics of coherence. In *ACL*, pages 391–398.

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.

Gideon Mann, Ryan McDonald, Mehryar Mohri, Nathan Silberman, and Dan Walker. 2009. Efficient large-scale distributed training of conditional maximum entropy models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1231–1239.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36, Los Angeles, California, June. Association for Computational Linguistics.

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

Neville Moray. 1959. Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 11(1):56–60.

Ani Nenkova and Kathleen McKeown. 2003. References to named entities: a corpus study. In *NAACL '03*, pages 70–72.

Malvina Nissim. 2006. Learning information status of discourse entities. In *Proceedings of EMNLP*, pages 94–102, Morristown, NJ, USA. Association for Computational Linguistics.

Jacki O'Neill and David Martin. 2003. Text chat in action. In *GROUP '03: Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work*, pages 40–49, New York, NY, USA. ACM Press.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195, Honolulu, Hawaii, October. Association for Computational Linguistics.

Massimo Poesio, Mijail Alexandrov-Kabadjov, Renata Vieira, Rodrigo Goulart, and Olga Uryupina. 2005. Does discourse-new detection help definite description resolution? In *Proceedings of the Sixth International Workshop on Computational Semantics*, Tillburg.

Amruta Purandare and Diane J. Litman. 2008. Analyzing dialog coherence using transition patterns in lexical and semantic features. In *FLAIRS Conference'08*, pages 195–200.

Dou Shen, Qiang Yang, Jian-Tao Sun, and Zheng Chen. 2006. Thread detection in dynamic text message

streams. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 35–42, New York, NY, USA. ACM.

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

Lidan Wang and Douglas W. Oard. 2009. Context-based message expansion for disentanglement of interleaved text conversations. In *Proceedings of NAACL-09*.