Neural Net Models for Open-Domain Discourse Coherence

Jiwei Li and Dan Jurafsky
Computer Science Department
Stanford University, Stanford, CA, 94305, USA
`jiweil,jurafsky@stanford.edu`

Abstract

Discourse coherence is strongly associated with text quality, making it important to natural language generation and understanding. Yet existing models of coherence focus on individual aspects of coherence (lexical overlap, rhetorical structure, entity centering) and are trained on narrow domains. We introduce algorithms that capture diverse kinds of coherence by learning to distinguish coherent from incoherent discourse from vast amounts of open-domain training data. We propose two models, one discriminative and one generative, both using LSTMs as the backbone. The discriminative model treats windows of sentences from original human-generated articles as coherent examples and windows generated by randomly replacing sentences as incoherent examples. The generative model is a SEQ2SEQ model that estimates the probability of generating a sentence given its contexts. Our models achieve state-of-the-art performance on multiple coherence evaluations. Qualitative analysis suggests that our generative model captures many aspects of coherence including lexical, temporal, causal, and entity-based coherence.[1]

1 Introduction

Modeling the discourse coherence of a text (the way parts of a text are linked into a coherent whole) is essential for tasks like summarization (Barzilay and McKeown, 2005), text planning (Hovy, 1988; Marcu, 1997) question-answering (Verberne et al., 2007), and even applications like psychiatric diagnosis (Elvevåg et al., 2007; Bedi et al., 2015).

Various frameworks exist, each tackling aspects of coherence. Lexical cohesion (Halliday and Hasan,

1976; Morris and Hirst, 1991) models chains of words and synonyms. Psychological models of discourse (Foltz et al., 1998; Foltz, 2007; McNamara et al., 2010) generalize lexical cohesion via LSA embeddings of sentences. Relational models like RST (Mann and Thompson, 1988; Lascarides and Asher, 1991) define relations that hierarchically structure texts. *Entity grid models* (Barzilay and Lapata, 2008) model the referential coherence of entities moving in and out of focus across a text. None of the models fully captures the rich semantic, discourse, and inferential links between coherent text units. Furthermore, previous work has been difficult to scale up and apply in open domains.

We propose to capture many of these aspects of discourse coherence (e.g., lexical, causal, entity focus) in neural net frameworks. We present two models: a discriminative model that induces coherence in an unsupervised manner from large datasets of real world texts by treating human generated texts as coherent examples and texts with random sentence replacements as negative examples; and a generative model that uses sequence-to-sequence models (Sutskever et al., 2014) (SEQ2SEQ) to model the likelihood of generating a sentence based on its context.

We evaluate the models on two text-ordering datasets, one from the literature (Barzilay and Lapata, 2008), and a new larger open-domain one. The discriminative model achieves state-of-the-art performance on the domain specific dataset presented in Barzilay and Lapata (2008), pushing the state-of-the-art result to 96% accuracy, significantly outperforming all previous models. The generative model obtains the best result on a large open-domain setting, including on the difficult task of reconstructing paragraph order, and qualitative evaluation suggests that it captures multiple types of coherence.

---

[1]System, code and datasets available upon publication.

2 Related Work

There are many frameworks for discourse coherence:

Lexical Coherence Coherence is strongly cued by words: words linked by identity, synonymy or other lexical relations forming chains across discourse segments (Halliday and Hasan, 1976). Early models used tools like thesauri (Morris and Hirst, 1991). Later work used Latent Semantic Analysis (LSA) embeddings (Foltz et al., 1998; Foltz, 2007), representing sentences with LSA vectors and measuring coherence with the cosine similarity of adjacent sentences, with the goal of capturing more subtle lexical relations that might not be available in thesauri.

Structured Discourse Relations Early work used discourse relations like Rhetorical Structure Theory (Mann and Thompson, 1988), a manually defined set of discourse relations between clauses, or Discourse Representation Theory (Lascarides and Asher, 1991)) a formal semantic model of discourse contexts, coreference and scope, to create coherent paragraphs in text planning (Hovy, 1988; Moore and Paris, 1989).

Entity Grid Models Many recent coherence models are based instead on *centering theory*, a model of which entity is in *focus* at a point in the discourse, and how smoothly that focus shifts from sentence to sentence depending, e.g., on the syntactic positions in which entities appear (Grosz et al., 1995; Walker et al., 1998; Strube and Hahn, 1999; Poesio et al., 2004). The most influential such model is the **entity grid** model of Barzilay and Lapata (2008), in which sentences are represented by a vector of coreferent discourse entities along with their grammatical roles. Probabilities of entity transitions between adjacent sentences are concatenated to document vector representation, used as the input to machine learning classifiers. Entity grid models have been extended with coreference (Elsner and Charniak, 2008), named entities (Eisner and Charniak, 2011), discourse relations (Lin et al., 2011), and entity graphs (Guinaudeau and Strube, 2013).

Neural Net Models Recent work focuses instead on representing sentences as dense, real-valued vectors (Ji and Eisenstein, 2014; Bhatia et al., 2015), such as by learning sentence representations as part of supervised RST discourse parsing (Li et al., 2014; Ji and Eisenstein, 2014).

Our proposed *discriminative* model extends the coherence model of Li and Hovy (2014), a neural classifier trained on small domain-specific datasets (earthquake and accidents) using negative sampling at the sentence level. The algorithm we present significantly outperforms the classifier of Li and Hovy (2014).

The proposed *generative* model uses a SEQ2SEQ backbone to generate a sentence from its contexts. SEQ2SEQ models have been successfully applied to a variety of NLP tasks including machine translation (Sutskever et al., 2014), dialogue generation (Vinyals and Le, 2015), and abstractive summarization (Rush et al., 2015). Our idea of predicting the current sentence based on the previous one is similar to skip-thought models (Kiros et al., 2015) that build an LSTM encoder-decoder model by predicting tokens in neighboring sentences. We use the mutual dependency between the two consecutive sequences to measure coherence. This idea of modeling the mutual dependency between two sequences for neural generation has been explored by Li et al. (2015) for dialogue generation.

The two models we propose can also be viewed as the a kind of generalization of the skip-gram model (Mikolov et al., 2013a; Mikolov et al., 2013b) to the sentence level. The generative model that predicts the next sentence based on the previous sentence is comparable to the skip-gram algorithm's predicting the next word given its context using a softmax function. The discriminative model is comparable to the negative sampling strategy (Goldberg and Levy, 2014) which has been widely used in training word embeddings.

3 Models

In this section, we describe the two models for coherence modeling, which are respectively suitable for different scenarios in real world applications.

3.1 The Discriminative Model

Notations Let C denote a sequence of coherent texts taken from original articles generated by humans. C is comprised of a sequence of sentences $C = \{s_{n-L}, ..., s_{n-1}, s_n, s_{n+1}, ..., s_{n+L}\}$ where L denotes the half size of the context window. Each sentence s is comprised of a sequence of words $s = \{w_1, w_2, ...\}$. Each word w is associated with
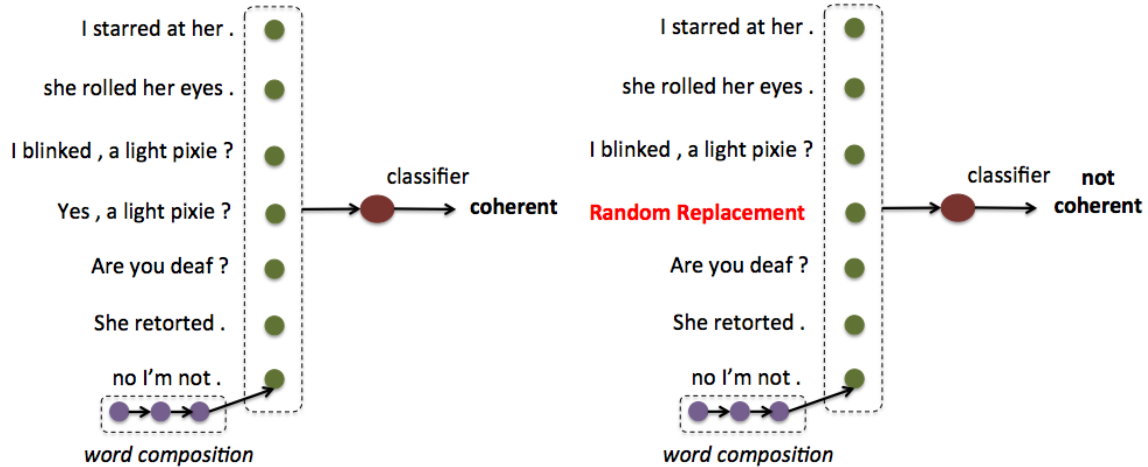
Figure 1: Illustrations of coherent (positive) vs not-coherent (negative) training examples for Concatenation Context Model.

a K dimensional vector h_w and each sentence is as well associated with a K dimensional vector x_s.

The model we propose is demonstrated in Figure 1. We treat cliques taken from the original articles as coherent positive examples and cliques with random replacements of center sentence s_n as negative ones. Each clique C is thus associated with a binary variable y_C indicating whether it is from original human generated articles or from random replacements.

Each clique C is associated with a $(2L + 1) \times K$ dimensional vector by concatenating the representations of its constituent sentences[2]. The sentence representation is obtained from LSTMs. For word compositions, we use the representation output from the last time step to represent the entire sentence. By concatenating representations of its constituent sentences, we obtain a $(2L + 1) \times K$ dimensional vector for C. We then map the $(2L + 1) \times K$ representation to a K dimensional vector using nonlinear composition:

$$L(C) = \tanh(W \cdot [x_{s_{n-L}}, ..., x_{s_n}, ..., x_{s_{n+L}}]) \quad (1)$$

where $W \in \mathbb{R}^{K \times (2L+1)K}$. To model negative incoherent examples, we resort to noise contrastive estimation (Gutmann and Hyvärinen, 2010). Let $p(y_C = 1|C)$ denote the possibility that clique C is coherent. Correspondingly, $p(y_C = 0|C) = 1 - p(y_C = 1|C)$

denotes the probability that the clique with sentence randomly sampled and incoherent. The probability that the current pair is coherent, i.e., $p(y_C = 1|C)$ is given by:

$$\log p(y_C = 1|C) = \log \frac{1}{1 + \exp(-U^T \cdot L(C))} \quad (2)$$

where $U \in \mathbb{R}^{1 \times K}$. For a given C, let C' denote the collection of negative cliques generated by replacing the middle sentence s_N. The loss function is then:

$$Loss = \log p(y_C = 1|C) + \sum_{C'} \log p(y_{C'} = 0|C') \quad (3)$$

The proposed model can be viewed as an extension of Li and Hovy's (2014) model but is practical at large scale[3].

Training Word vectors and LSTM parameters are randomly initialized from the uniform distribution [-0.1,0.1]. Since the model does not require softmax for word prediction, we keep a relative large vocabulary of the top 200,000 most frequent words. We adopt stochastic gradient decent with min-batch size 128 and clip the gradients if the norm of gradient vectors exceed 5. We set the number of negative examples to 10, 5 of which are sampled from the same document, and the rest from random documents. We use a dropout rate of 0.2 in training and employ no

---

[2]To classify first/last sentences and include their context, we would need special beginning/ending sentence vectors; for simplicity, we treat all elements in beginning/ending sentence vectors as zero.

[3]Li and Hovy's (2014) recursive neural model operates on parse trees, which does not support batched computation and is therefore hard to scale up.

regularizers. We run 7 epochs with initial learning rate of 1.0. After 4 iterations, we begin halving the learning rate after each epoch.

3.2 The Generative Model

In a coherent context, a machine should be able to guess the next utterance given the previous one. We therefore propose measuring the degree of coherence using the likelihood of observing a sentence given its context.

Given two consecutive sentences $[s_i, s_{i+1}]$, we measure the coherence by combining the likelihood of generating s_i given s_{i+1} and generating s_{i+1} given s_i:

$$L(s_i, s_{i+1}) = \frac{1}{2}[\log p(s_i|s_{i+1}) + \log p(s_{i+1}|s_i)] \quad (4)$$

Eq.4 measures the mutual dependency between the two consecutive sentences. Both $p(s_i|s_{i+1})$ and $p(s_{i+1}|s_i)$ can be computed using SEQ2SEQ models (Sutskever et al., 2014). SEQ2SEQ models define a distribution over outputs y and sequentially predict tokens using a softmax function:

$$p(y|x) = \prod_{t=1}^{n_T} \frac{\exp(f(h_{t-1}, e_{y_t}))}{\sum_{w'} \exp(f(h_{t-1}, e_{w'}))}$$

where $f(h_{t-1}, e_{y_t})$ denotes the activation function between h_{t-1} and e_{w_t}, where h_{t-1} is the representation output from the LSTM at time $t-1$. Each sentence concludes with a special end-of-sentence symbol *EOS*. Commonly, the input and output each use different LSTMs with separate sets of compositional parameters to capture different compositional patterns. During decoding, the algorithm terminates when an *EOS* token is predicted.

We separately train two models: $p(s_{i+1}|s_i)$ that predicts the next sentence based on the previous one from the original passages and and $p(s_i|s_{i+1})$ that predicts the previous sentence given the next sentence. $p(s_i|s_{i+1})$ can be trained in the similar way as $p(s_{i+1}|s_i)$ with sources and targets swapped. To avoid the model favoring shorter sequences, the log likelihood is divided by the length of the sequence.

Training We adopt a deep structure with four LSTM layers for encoding and four LSTM layers for decoding, each of which consists of a different set of parameters. Each LSTM layer consists of 1,000

Dataset	words	sentences	words per sent.
Wikipedia	2.1B	142M	14.8

Table 1: Statistics for the Datasets.

hidden neurons, and the dimensionality of word embeddings is set to 1,000. We keep a vocabulary of the top 50,000 most frequent words. Other training details are given below, broadly aligned with Sutskever et al. (2014): LSTM parameters and embeddings are initialized from a uniform distribution in [-0.1, 0.1]; Stochastic gradient decent is implemented using a fixed learning rate of 0.1; Batch size is set to 128; Gradient clipping is adopted by scaling gradients when the norm exceeded a threshold of 5. We run 8 epochs in total and start halving the learning rate after 5 epochs.

4 Experimental Results

In this section, we describe experimental results. We first evaluate the proposed model on the task of sentence ordering using two datasets, a standard domain-specific dataset (Barzilay and Lapata, 2008) and a newly constructed open-domain dataset from Wikipedia. Next we propose the task of paragraph reconstruction that reconstruct an original paragraph from its constituent sentences whose order has been permuted.

4.1 Sentence Ordering, Domain-specific Data

Dataset We first evaluate the proposed algorithms on a dataset widely adopted in sentence ordering and predicate on the assumption that an article is always more coherent than a random permutation of its sentences (Barzilay and Lapata, 2008; Louis and Nenkova, 2012; Elsner et al., 2007; Lin et al., 2011). The corpus consists of 200 articles each from two domains: NTSB airplane accident reports (V=4758, 10.6 sentences/document) and AP earthquake reports (V=3287, 11.5 sentences/document), split into training and testing. For each document, pairs of permutations are generated[4]. Each pair contains the original document order and a random permutation of the sentences from the same document.

We use reduced versions of both our models to

---

[4] Permutations downloaded from `people.csail.mit. edu/regina/coherence/CLsubmission/`.

Model	Accident	Earthquake	Average
Our Discriminative Model	**0.930**	**0.992**	**0.956**
Our Generative Model	0.755	0.930	0.842
Recursive Neural Models (Li et al. 2014)	0.864	0.976	0.920
Entity Grid Model (Barzilay and Lapata, 2008)	0.904	0.872	0.888
HMM (Louis and Nenkova, 2012)	0.822	0.938	0.880
HMM+Entity (Louis and Nenkova, 2012)	0.842	0.911	0.876
HMM+Content (Louis and Nenkova, 2012)	0.742	0.953	0.847
Graph (Guinaudeau and Strube, 2013)	0.846	0.635	0.740
Foltz et al. (1998)-Glove	0.705	0.682	0.693
Foltz et al. (1998)-LDA	0.660	0.667	0.664

Table 2: Results from different coherence models. Baseline numbers from prior work (except for Foltz et al. (1998)) are reprinted from the best performance reported in those papers.

allow fair comparison with baselines. We therefore do not use the massive Wikipedia training set, holding the datasize constant and training only on the earthquake/traffic training set. For the discriminative model, we generate noise negative examples from random replacements in the training set, with the only difference that random replacements only come from the same document. We use 300 dimensional embeddings borrowed from GLOVE (Pennington et al., 2014) to initialize word embeddings. Word embeddings are kept fixed during training and we update LSTM parameters using AdaGrad (Duchi et al., 2011). For the generative model, due to the small size of the dataset, we train a one layer LSTM SEQ2SEQ model with word dimensionality and number of hidden neurons set to 100.

We report performances from the following widely used baselines in coherence literature.

(1) *Entity Grid Model*: The grid model (Barzilay and Lapata, 2008) represents the sentence as a column of a grid of features and applies machine learning methods (e.g., SVM) to identify the coherent transitions based on entity features. Results are directly taken from Barzilay and Lapata's (2008) paper.

(2) *HMM*: A hidden-markov model described in Louis and Nenkova (2012) models the cluster transition probability in the coherent texts. Results are from their paper.

(3) *Graph Based Approach*: Guinaudeau and Strube (2013) extended the entity grid model to a graph representing the text that embeds entity transition information needed for local coherence computation (Guinaudeau and Strube, 2013).

(4) Li and Hovy (2014): A recursive neural model computes sentence representations based on parse trees. Negative sampling is used to construct negative incoherent examples. Representations of neighboring sentences are concatenated and fed into a neural classification, outputting whether a clique of sequences is coherent or not. Results are from their paper

(5) Foltz et al. (1998) computes the semantic relatedness of two text units as the cosine similarity between their LSA vectors. The coherence of a discourse is the average of the cosine of adjacent sentences. We used this intuition, but with more modern embedding models: (1) 300-dimensional Glove word vectors (Pennington et al., 2014), embeddings for a sentence computed by averaging the embeddings of its words (2) Sentence representations obtained from LDA (Blei et al., 2003) with 300 topics, trained on the Wikipedia dataset using Gibbs sampling. We compute coherence as the average cosine between adjacent sentences. Since embeddings are pre-trained, these models do not make use of training data.

Results are reported in Table 2. The proposed discriminative model significantly outperforms the model presented in Li and Hovy (2014) as well as all non-neural baselines. It achieves roughly 100% accuracy on the earthquake dataset and 93% on the accident dataset, marking a significant advancement in the benchmark. The generative model does not perform competitively on this dataset. This is due to the small size of the dataset, leading the generative model to overfit.

The simple LSA method of calculating cosine similarity between adjacent sentences, adopted from Foltz et al. (1998), does not yield competitive results, con-

firming that while simple centroids of word embeddings may do a good job of modeling lexical coherence, lexical coherence is only one component of discourse coherence.

4.2 Evaluating Ordering on Open-domain

Since the dataset presented in Barzilay and Lapata (2008) is quite domain-specific, we propose testing coherence with a much larger, open-domain dataset: Wikipedia. We created a test set by randomly selecting 984 paragraphs from Wikipedia dump 2014, each paragraph consisting of at least 16 sentences. The training set is the 80 million sentences. We ensure that there is no overlap between the training set and the test set. Based on this dataset, we define the following tasks for evaluation:

4.2.1 Binary Permutation Classification

We adopt the same strategy as in Barzilay and Lapata (2008), in which we generate permutations for the original Wikipedia paragraphs. We follow the protocols described in the subsection above to compare the degree of coherence between the original texts and their permutations. Each pair whose original paragraph's score is higher than its permutation is treated as being correctly classified, else incorrectly classified. Models are evaluated using accuracy.

Baselines   Our baselines consist of the Glove and LDA updates of the lexical coherence baselines (Foltz et al., 1998). We also implement the Entity Grid Model (Barzilay and Lapata, 2008) using the Wikipedia training set. For each noun in a sentence, we extract its syntactic role (subject, object or other). We use a wikipedia dump parsed using the Fanse Parser (Tratz and Hovy, 2011). Subjects and objects are extracted based on *nsubj* and *dobj* relations in the dependency trees. (Barzilay and Lapata, 2008) define two versions of the Entity Grid Model, one using full coreference and a simpler method using only exact-string coreference; Due to the difficulty of running full coreference resolution over 80 million Wikipedia sentences, we follow other researchers in using Barzilay and Lapata's simpler method (Feng and Hirst, 2012; Burstein et al., 2010; Barzilay and Lapata, 2008).[5] We also employ the *uni-directional*

---

[5]Our implementation of the Entity Grid Model is built upon public available code at `https://github.com/karins/CoherenceFramework`.

Model	Accuracy
Our Generative Model (bi-directional)	**0.886**
Our Generative Model (uni-directional)	0.842
Our Discriminative Model	0.715
Entity Grid Model	0.686
Foltz et al. (1998)-Glove	0.597
Foltz et al. (1998)-LDA	0.575

Table 3: Performance on the open-domain binary classification dataset of 984 Wikipedia paragraphs.

baseline in which the coherence score is computed using only $p(s_{i+1}|s_i)$, i.e., predicting the next sentence given the previous one.

Results   Figure 3 presents results on the binary classification task. Once again, purely lexical methods (Foltz et al., 1998) do not yield compelling results. Contrary to the findings on the domain specific dataset in the previous subsection, the discriminative model does not yield compelling results, performing only slightly better than the entity grid model. We believe the poor performance is due to the sentence-level negative sampling used by the discriminative model. Due to the huge semantic space in the open-domain setting, the sampled instances can only cover a tiny proportion of the possible negative candidates, and therefore don't cover the space of possible meanings. By contrast the dataset in Barzilay and Lapata (2008) is very domain-specific, and the semantic space is thus relatively small. By treating all other sentences in the document as negative, the discriminative strategy's negative samples form a much larger proportion of the semantic space, leading to good performance.

The proposed generative model performs significantly better than all other baselines. Compared with the dataset in Barzilay and Lapata (2008), overfitting is not an issue here due to the great amount of training data. In line with our expectation the bi-directional model which models the bidirectional dependency between the two consecutive sentences outperforms the uni-directional model which only handles the case of predicting the next sentence.

4.2.2 Paragraph Reconstruction

The accuracy of our models on the binary task of detecting the original sentence ordering is very high, on both the prior small task and our large open-domain version. We therefore believe it is time for

First Sent Being Given	
Our Generative Model (bi-directional)	**0.216**
Our Generative Model (uni-directional)	0.208
Entity Grid Model	0.153
Lexical Model (Glove)	0.084
No Clue Being Given	
Our Generative Model	**0.175**
Entity Grid Model	0.124

Table 4: Performances of the proposed models on the open-domain paragraph reconstruction dataset.

the community to move to a more difficult task for measuring coherence.

We suggest the task of *reconstructing an original paragraph from a bag of constituent sentences*, which has been previously used in coherence evaluation (Lapata, 2003). More formally, given a set of permuted sentences $s_1, s_2, ..., s_N$ (N the number of sentences in the original document), our goal is return the original (presumably most coherent) ordering of s.

Because the *discriminative model* calculates the coherence of a sentence given the known previous and following sentences, it cannot be applied to this task since we don't know the surrounding context. Hence, we only use the *generative model*. We explore the following two settings:

(1) The first sentence of a paragraph is given: for each step, we compute the coherence score of placing each remaining candidate sentence to the right of the partially constructed document. We use beam search with beam size 10.[6]

(2) No clue is given: we employed the graph based method described in Lapata (2003). We first construct a graph where the each vertex denotes a sentence and the edge weight $u \rightarrow v$ denotes the coherence score of sentence v coming after u. Note that weight values for $u \rightarrow v$ and $v \rightarrow u$ are different. We initialize the vertex list V using all vertexes in the graph. Similar to Lapata (2003), we employ a greedy search model. The greedy algorithm first picks the edge $u \rightarrow v$ with the highest coherence score, and deletes all the outgoing edges from vertex u and all incoming edges to vertex v. u, v are removed from the vertex list V. Next, for each time step, let v_{left} and v_{right} respectively denote the left-most

---

[6]Not guaranteed to find the optimal solution since this generation task is known to be NP-complete (0).

and right-most node in the partially constructed paragraph. The greedy model chooses whether to expand the paragraph to the left and to the right by comparing $\max_{v' \in V} \text{S}(v', v_{\text{left}})$ with $\max_{v' \in V} \text{S}(v_{\text{right}}, v')$, where the former denotes the maximal coherence score of placing a remaining sentence to the left of the partially constructed paragraph and the latter denotes the maximal score of appending a sentence to the right of the paragraph. The newly selected node is added to the paragraph and removed from the vertex list V. We repeat this process until V is empty.

We use the Entity Grid model as a baseline for both the settings. LSA-style cosine similarity based lexical methods are symmetric regarding the next sentence and the previous sentence. We therefore can not tell which sentence should come first. We thus only use it as a baseline in the first-sentence-being-given setting.

Evaluating the absolute positions of sentences would be too harsh, penalizing orderings that maintain relative position between sentences through which local coherence can be manifested. We therefore use Kendall's Tau (Lapata, 2003; Lapata, 2006), a metric of rank correlation for evaluation. Kendall's τ is computed based on the number of inversions in the rankings as follows:

$$\tau = 1 - \frac{2\# \text{ of inversions}}{N \times (N - 1)} \quad (5)$$

where N denotes the number of sentences in the original document and inversions denote the number of interchanges of consecutive elements needed to reconstruct the original document. Kendall's τ can be efficiently computed by counting the number of intersections of lines when aligning the original document and the generated document. We refer the readers to Lapata (2003) for more details.

Results are reported in Figure 4. The generative model outputs both the Entity Grid model and the lexical model by a large margin. In line with our expectations, better scores are observed in the first-sentence-given setting than the no-clue-given setting, We again observe a performance boost from the bi-directional model over the uni-directional model.

4.3 Qualitative Analysis

To investigate which kinds of coherence the model is capable of handling, we examine some relevant examples, annotated with the (log-likelihood) coherence

score from the generative model. Each of the examples above/below was chosen in advance, before we trained our model, hence were not "cherry-picked".

Case 1: Lexical Coherence
Pinochet was arrested. His arrest was unexpected. **-4.25**
Pinochet was arrested. His death was unexpected. **-4.68**
Mary ate some apples. She likes apples. **-5.66**
Mary ate some apples. She likes pears. **-6.16**
Mary ate some apples. She likes Paris. **-6.72**

The model can handle lexical coherence, correctly favoring the 1st over the 2nd, and the 3rd over the 4th examples.

Case 2: Temporal Order

Washington was unanimously elected president in the first two national elections. He oversaw the creation of a strong, well-financed national government. **-3.48**
Washington oversaw the creation of a strong, well-financed national government. He was unanimously elected president in the first two national elections. **-4.52**

Case 3: Causal Relationship
Bret enjoys video games; therefore, he sometimes is late to appointments. **-7.59**
Bret sometimes is late to appointments; therefore, he enjoys video games. **-7.64**

The model also does well at the much more complex task of dealing with temporal and causal relationships. From its training the model is exposed to the general preference of natural text for temporal order, and even for the more subtle causal links.

Case 4: Centering/Referential Coherence
Mary ate some apples. She likes apples. **-5.66**
She ate some apples. Marry likes apples. **-7.64**

The model can handle simple cases of referential coherence.
Example3: **-3.72**
John went to his favorite music store to buy a piano.
He had frequented the store for many years.
He was excited that he could finally buy a piano.
He arrived just as the store was closing for the day.
Example4: **-4.55**
John went to his favorite music store to buy a piano.
It was a store John had frequented for many years

He was excited that he could finally buy a piano..
It was closing just as John arrived.

In these examples from Miltsakaki and Kukich (2004), the model successfully captures the fact that the second text is less coherent due to *rough shifts*. The model, in mapping sentences to semantic vector space successfully captures a representation of entity focus and its subtle syntactic cues.

5 Conclusion

We investigate the problem of discourse coherence, treating natural texts as coherent and permutations as non-coherent, and training large neural models that achieve state of the art performance on coherence, including on large open-domain test sets. The performance and our qualitative analysis suggest that the distributed sentence representations built by the model capture some of the implicit linguistic components of coherence. Our model outperforms LSA baselines, suggesting it models lexical coherence well, and seems to capture semantic coherence like temporal and causal relations, which prior models like LSA and grid-based models are not designed to capture. The model also outperforms grid-based models, suggesting it may do well at capturing coherence based on entity focus across a discourse.

SEQ2SEQ models have achieved recent success in many generation tasks. The fact that our generative model does well at the open-domain tasks, including paragraph reconstruction despite its known difficulty (0), suggests that SEQ2SEQ models can also play an important role for modeling discourse coherence.

6 Acknowledgement

References

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

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Regina Barzilay and Kathleen R McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, pages 681–684. Association for Computational Linguistics.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Micha Eisner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 125–129. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 41–44. Association for Computational Linguistics.

Micha Elsner, Joseph L Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *HLT-NAACL*, pages 436–443.

Brita Elvevåg, Peter W Foltz, Daniel R Weinberger, and Terry E Goldberg. 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1):304–316.

Vanessa Wei Feng and Graeme Hirst. 2012. Extending the entity-based coherence model with multiple ranks. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 315–324. Association for Computational Linguistics.

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

Peter W Foltz. 2007. Discourse coherence and lsa. *Handbook of latent semantic analysis*, pages 167–184.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

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

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 93–103.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, pages 297–304.

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

Eduard H Hovy. 1988. Planning coherent multisentential text. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*, pages 163–169. Association for Computational Linguistics.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 13–24.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3276–3284.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 545–552. Association for Computational Linguistics.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

Alex Lascarides and Nicholas Asher. 1991. Discourse relations and defeasible knowledge. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 55–62. Association for Computational Linguistics.

Jiwei Li and Eduard Hovy. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 997–1006. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2012. A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1157–1168. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu. 1997. From local to global coherence: A bottom-up approach to text planning. In *AAAI/IAAI*, pages 629–635. Citeseer.

Danielle S. McNamara, Max M. Louwerse, Philip M. McCarthy, and Arthur C. Graesser. 2010. Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(01):25–55.

Johanna D Moore and Cecile L Paris. 1989. Planning text for advisory dialogues. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pages 203–211. Association for Computational Linguistics.

J. Morris and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

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

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

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

Ilya Sutskever, Oriol Vinyals, and Quoc Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Stephen Tratz and Eduard Hovy. 2011. A fast, accurate, non-projective, semantically-enriched parser. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1257–1268. Association for Computational Linguistics.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–736. ACM.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Marilyn A Walker, Aravind Krishna Joshi, and Ellen Friedman Prince. 1998. *Centering theory in discourse*. Oxford University Press.