How coherent are neural models of coherence?

Leila Pishdad Federico Fancellu Ran Zhang Afsaneh Fazly
Samsung AI Centre Toronto (SAIC Toronto)
{leila.p, federico.f, ran.zhang, a.fazly}@samsung.com

Abstract

Despite the recent advances in coherence modelling, most such models including state-of-the-art neural ones, are evaluated on either contrived proxy tasks such as the standard order discrimination benchmark, or tasks that require special expert annotation. Moreover, most evaluations are conducted on small newswire corpora. To address these shortcomings, in this paper we propose four generic evaluation tasks that draw on different aspects of coherence at both the lexical and document levels, and can be applied to any corpora. In designing these tasks, we aim at capturing coherence-specific properties, such as the correct use of discourse connectives, lexical cohesion, as well as the overall temporal and causal consistency among events and participants in a story. Importantly, our proposed tasks either rely on automatically-generated data, or data annotated for other purposes, hence alleviating the need for annotation specifically targeted to the task of coherence modelling. We perform experiments with several existing state-of-the-art neural models of coherence on these tasks, across large corpora from different domains, including newswire, dialogue, as well as narrative and instructional text. Our findings point to a strong need for revisiting the common practices in the development and evaluation of coherence models.

1 Introduction

A variety of research has focused on modelling *textual coherence*, with the goal of distinguishing coherent from incoherent documents. This distinction has important repercussions for a variety of downstream tasks, including automatic essay scoring, document summarization, and natural language generation (Jernite et al., 2017; Li and Jurafsky, 2017; Wu and Hu, 2018). Given the importance of the task, there is a long line of approaches proposed for modelling coherence.

Most earlier models leverage linguistic features to discriminate between coherent and incoherent documents: The classic work of Barzilay and Lapata (2008), and several follow-up studies (Elsner and Charniak, 2011; Guinaudeau and Strube, 2013; Zhang et al., 2015) model coherence by using entity transitions — information about how semantic roles change across sentences. Lin et al. (2011) and Feng et al. (2014) rely on discourse relation transitions, whereas Louis and Nenkova (2012) focus on patterns of syntactic transition extracted from a constituency tree. Alternatively, other models have drawn on topic transitions across sentences in a document to assess its coherence (Morris and Hirst, 1991; Somasundaran et al., 2014; Mesgar and Strube, 2016).

Recent models on the other hand, rely on neural architectures with little to no feature engineering required. Nguyen and Joty (2017) and Mohiuddin et al. (2018) propose neural variants of the entity transition model of Barzilay and Lapata (2008), while others devise approaches to learning distributed representations of sentences/documents that also encode coherence (Li and Hovy, 2014; Jernite et al., 2017; Li and Jurafsky, 2017; Logeswaran et al., 2018). These specialized encoders often claim to combine various aspects of coherence into a unified framework (Mesgar and Strube, 2018; Mim et al., 2019; Moon et al., 2019). However, despite the recent advances in coherence modeling, most approaches are still

Proceedings of the 28th International Conference on Computational Linguistics, pages 6126–6138
Barcelona, Spain (Online), December 8-13, 2020

evaluated using a benchmark known as the *order discrimination* task, where the goal is to discriminate between an original coherent document and its variants in which sentences are randomly shuffled.

While this method allows to train and evaluate models of coherence without the need for expensive human annotations, an important question left unanswered is what aspects of coherence this evaluation captures. High performance on such a proxy task does not necessarily mean that a model is capable of identifying *naturally*-incoherent documents, as noted by many researchers, and empirically demonstrated in a recent study by Lai and Tetreault (2018). Despite this important limitation, a majority of coherence models focus on optimizing performance on the order discrimination task, and mainly on news corpora (such as the Wall Street Journal). Although such standardization of task and data makes comparison across models easy, it also undesirably encourages the community to gradually move away from the real goal of modelling coherence, and towards optimizing performance on the specific test itself.

To address the above concerns, we propose four new tasks for evaluating different aspects of coherence, including appropriate use of connectives, consistency in topics, as well as overall temporal and causal cohesion among events and participants. We perform extensive evaluation of several existing state-of-the-art (SOTA) models on these tasks, across a variety of corpora from different domains, including news, dialogue, as well as narrative and instructional text. Importantly, our proposed evaluation tasks do not require specialized human annotations, and instead levarage automatic methods or pre-existing annotated corpora. Similar to the standard order discrimination task, our proposed tasks also aim at distinguishing an original coherent document from an incoherent variant generated by manipulating certain elements of the original document. Our results show a complex landscape of model performances that vary across tasks and corpora.

2 Background

Models of coherence can be divided into two main general approaches, with implications on how these models are evaluated. One group is concerned with directly measuring human ratings of coherence and/or text quality, with applications in language learning and automatic essay scoring (Pitler and Nenkova, 2008; Somasundaran et al., 2014; Clercq and Hoste, 2016; Mesgar and Strube, 2016; Mesgar and Strube, 2018; Lai and Tetreault, 2018; Mim et al., 2019). A second group of studies focuses on modeling coherence for the purpose of learning better representations that can improve downstream natural language processing applications, such as summarization and generation (Barzilay and Lapata, 2008; Elsner and Charniak, 2011; Lin et al., 2011; Louis and Nenkova, 2012; Guinaudeau and Strube, 2013; Feng et al., 2014; Li and Hovy, 2014; Zhang et al., 2015; Li and Jurafsky, 2017; Nguyen and Joty, 2017; Jernite et al., 2017; Wu and Hu, 2018; Mohiuddin et al., 2018; Logeswaran et al., 2018; Moon et al., 2019). The former group evaluates models by examining their correlation with human ratings of coherence and readability, whereas the latter develops proxies of coherence, such as the sentence order discrimination task. This second group is the most relevant to our study, and as such we further elaborate on several proxy tasks that have been commonly used for evaluating and comparing models of coherence.

By far the most common evaluation technique is the order discrimination task, where a coherence model is required to distinguish between an original document and a set of incoherent variants automatically generated by randomly shuffling the sentences in the original text. Earlier studies simply assumed that a shuffled variant of a document is less coherent than the original, an assumption later verified by Lin et al. (2011). This task was initially proposed by Barzilay and Lapata (2008), and has since become a standard benchmark for developing and evaluating coherence models. However, as noted by Barzilay and Lapata (2008), "the synthetic data used in the [order discrimination] task only partially approximates coherence violations that human readers encounter". Others have also noted that the order discrimination may be much easier than the actual task of coherence rating, hence proposing improved variants of this task. For example, Elsner and Charniak (2011) suggest the *insertion* task, which evaluates a coherence model based on how well it can predict the original position of each sentence within a document. Moon et al. (2019) evaluate their model on a more challenging *local discrimination* task, where the random shuffling is performed locally (e.g., within a 3-sentence window) to generate incoherent variants of an original document. Jernite et al. (2017) propose *next sentence prediction* that aims to predict which of the

next 5 sentences in a document is the correct continuation given a history/context — a task they use as a discourse-based objective to learn text representations that also encode coherence.

Recognizing the limitations of the order discrimination task, many evaluate their models on additional tasks that aim at rating the overall quality of text, including *summary coherence rating*, *readability assessment*, and *essay scoring* (Barzilay and Lapata, 2008; Guinaudeau and Strube, 2013; Feng et al., 2014; Li and Hovy, 2014; Zhang et al., 2015; Nguyen and Joty, 2017). In the summary coherence rating task, the goal is to rank a human-generated document summary higher than a machine-generated summary of an original document, whereas for readability assessment a model must discriminate between an original difficult-to-read document and its easy-to-read simplified variant. Although such evaluation tasks complement the order discrimination task and its variants, they still lack full interpretability since the overall quality and/or readability of a document depends on many factors of which coherence is one. Lai and Tetreault (2018) address this limitation by compiling a new annotated dataset to evaluate coherence "in the wild". Their dataset contains "real-world" user-generated text, such as emails and online reviews, each annotated by experts to reflect its level of coherence. We take a significant step further from the above work and propose four new evaluation tasks that each capture a different aspect of coherence, without requiring coherence-specific annotations.

3 Our proposed tasks

We propose four tasks that capture various aspects of coherence, by manipulating a text at the local lexical and global document levels. At the lexical level, we focus on the role of discourse connectives (*connective substitution*), and overall lexical cohesion (*sentence cloze*). At the document level, we assess whether coherence models can detect a sudden shift in topic (*topic switching*), and whether they can predict the correct last sentence of a document to create a coherent story (*story cloze*).

3.1 Connective substitution

Discourse connectives are linguistic expressions (words or phrases) that signal semantic relations, such as condition or result, between two text units; and are known to significantly contribute to text coherence and readability (Halliday and Hasan, 1976). Our goal is to automatically generate incoherent variants of a document (negative samples) via manipulating the connectives. We simulate incoherence in the use of connectives by replacing them with other connectives that are a good fit within the sentence context, but carry a different discourse relation (across sentences) then the original ones. For instance in (1), the connective 'at the same time' which signals that two events happen concurrently is substituted by 'thereafter', which instead means that one event follows another.

(1) a. *At the same time*, six ANC colleagues, [...] were reunited with their families
    b. *Thereafter*, six ANC colleagues, [...] were reunited with their families

Our task requires that we know which discourse connectives to target, and what meaning they convey. We thus use the Penn Discourse Treebank (PDTB; Prasad et al. (2008)),[1] in which each connective is tagged with a main and a secondary semantic category. We propose four strategies for generating negative samples, based on the cross-combination of the following conditions:

- In the *different* condition, we ensure that the semantic categories of the original and the substituted connectives are different. In the *same* condition, the two come from the same category, but a different subcategory (to allow for some degree of discourse meaning shift).
- In the *full* condition, we modify all sentences in a document that contain a connective, replacing one connective at random if a sentence contains more than one. In the *half* condition, we replace 50% of the sentences at random.

Given an original sentence containing a connective, we need to ensure that the replacement, i.e., the substitute connective, results in a natural English sentence. For this purpose, we use BERT (Devlin

[1] https://www.seas.upenn.edu/~pdtb/

et al., 2018).[2] Specifically, we mask the connective in the original sentence, and select our candidate replacements from the top-20 predictions output by BERT.[3] In the absence of human judgement, this procedure is a proxy to ensure that individual sentences in the new document are syntactically and semantically valid, while the overall coherence of the document has changed due to modifying the semantics of the discourse relations conveyed through connectives.

3.2 Sentence cloze

Our sentence cloze task is inspired by language proficiency tests, such as the cloze test, and the multiple-choice reading comprehension tests for language learners. Variants of the cloze task (Taylor, 1953; Deyes, 1984; Chambers and Jurafsky, 2008) assess language proficiency in a system (or human) by removing a random word from a sentence, and asking the system/human to fill in the blank, e.g., *Today, I went to the ... to send a letter*. These tests can evaluate syntactic, semantic, or discourse knowledge, depending on the type of the word that is removed. Inspired by the approach of Susanti et al. (2018), we generate incoherent variants of a document by replacing a randomly-selected word (noun or verb) in every sentence in the original text with an automatically-generated *distractor*. The distractors are chosen such that they fit in the sentence, but have a different meaning, e.g., we may replace *sold* for *bought*. Replacing a verb/noun in each sentence of a document is expected to change the overall discourse structure of the document, by changing the temporal and/or causal relations among the events and entities. To generate distractors, we first sample candidates amongst antonyms, hypernyms and hyponyms of the original word using WordNet, to ensure that the meaning of the distractor is semantically related to the original word. We then select the top 5 candidates by ranking the sentences with the distractors using the GPT-2 model (Radford et al., 2019).[4] We expect this process to generate less coherent documents, by lowering the *lexical cohesion* in the original documents, which is defined as the overall connectedness arising from semantic relationships between words across sentences (Halliday and Hasan, 1976; Morris and Hirst, 1991).

Similar to the connective substitution task, we experiment with two different conditions: a) The *full* substitution condition, where we substitute one noun/verb for every sentence in a document, and b) The *half* condition where we substitute one word in only 50% of the sentences chosen at random.

3.3 Topic switching

A major issue with the standard order discrimination task is that the automatically-generated incoherent documents (with random sentence shuffling) would have very low readability because they lack any topic structure. We would like to test coherence models on their ability to detect a *shift in topics*, as opposed to a complete loss of structure. Thus, we aim to generate variants of an original document that preserve some of the topic structure, but introduce a shift in topics half-way through the document. To achieve this goal, we glue together one half of an original document with a second half from another document. We propose two variants of the topic switching:

- The *controlled* mixing approach that ensures little overlap between the topics of the two documents to be mixed, based on topics automatically learned using the Latent Dirichlet Allocation (LDA) method of Blei et al. (2003).[5]

- The *random* mixing approach randomly mixes two documents, without explicitly controlling for degree of topic overlap between the documents. In order for the models not to be biased by the order of the halves, we randomly select if it is the first or the second half to be substituted.

3.4 Story cloze

Story cloze is a task originally proposed by Mostafazadeh et al. (2016) for story understanding and generation, with the goal of capturing temporal and causal relations between events as a central component

[2]Available through the HuggingFace Transformers library (Wolf et al., 2019) at `https://huggingface.co/`

[3]The average rank of the candidate replacement is 5.53.

[4]Similar to BERT, this is also available through the HuggingFace Transformers library.

[5]In order to obtain meaningful topics and to get rid of noisy high frequency words, we only consider the top $n\%$ vocabulary ranked using TF-IDF, where n is determined empirically for each corpus.

of coherence. Importantly, Mostafazadeh et al. (2016) also create a collection of non-fictional short stories, noting that the key property in the collection of these stories has been that "the story should read like a coherent story". We propose to use two variants of the story cloze test for the evaluation of coherence models. In the *original* story cloze test, each story is comprised of a short context (one or a few sentences) followed by two or more possible endings. A model is asked to choose between the correct ending that presents a coherent continuation of the context, and incorrect endings (or distractors) that are relevant in terms of topic, action and participants involved but would lead to an incoherent story. See example (2) where b) is a plausible story ending for a), but c) contradicts it, although the participants are the same.

(2) a. Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.
b. Karen became good friends with her roommate.
c. Karen hated her roommate.

4 Selected coherence models

We select four existing neural coherence models to carry out evaluations on, based on two criteria: performance-wise, they were shown to yield good performance on the order discrimination task; model-wise, we wanted models that differ in their underlying architecture to assess whether this has an impact on the different coherence tasks. Specifically, we focus on models that build upon and are extensions of *raw text encoders*; this choice is mainly dictated by the nature of some of our tasks where we require a model to make a local choice (at the word or sentence level) depending on the surrounding context. We do not consider models that are based on entity grids since they require an added level of abstraction (i.e., grammatical roles extracted from constituency trees), and encode context as a simple chain of noun-specific grammatical roles. We briefly describe the four models next.

Sentence averaging (SentAvg). SentAvg is a simple model proposed by Lai and Tetreault (2018) that learns a document representation by averaging over sentence embeddings, which are in turn learned using a single-layer LSTM (Hochreiter and Schmidhuber, 1997) over pre-trained GloVe (Pennington et al., 2014) word embeddings. The document-level embedding vectors are then passed through a linear transformation and a softmax layer to predict each document's coherence level.

Paragraph sequence (ParSeq). SentAvg is a simple model that ignores structure of documents. To address this limitation, Lai and Tetreault (2018) proposed ParSeq that takes into account paragraph structure within a document, as well as the order of sentences in each paragraph. ParSeq extends SentAvg to include three stacked LSTMs: a sentence LSTM layer first learns sentence embeddings from input GloVe word embeddings; a paragraph LSTM layer then takes these sentence embeddings and outputs a paragraph vector; and finally, a document LSTM layer takes these paragraph embeddings and learns a document-level vector. Similar to SentAvg, a linear transformation and a softmax layer are then applied to predict the coherence of a document.

Neural local coherence model (CohLSTM). Similar to SentAvg and ParSeq, the CohLSTM model of Mesgar and Strube (2018) learns sentence embeddings using a layer of LSTM over pre-trained word embeddings. However, CohLSTM captures the salient semantic information that connects two consecutive sentences by taking the average of their most similar LSTM hidden states, where similarity is captured by the dot-product of the corresponding vectors. These average vectors are then passed through a function that takes the pair-wise similarity scores of consecutive vectors (representing consecutive sentence pairs) normalized by the input length. Through the application of convolution on these vectors, Mesgar and Strube (2018) extract and represent patterns of semantic (topic) changes in a document, which are then used to calculate an overall coherence score.

Unified neural coherence model (UNC). The UNC model of Moon et al. (2019) aims to capture both local and global aspects of coherence, where the former refers to the connection between adjacent sentences, and the latter deals with coherence at the document level. Their model uses a Siamese

	VIZSIS	ROCSTORIES	SELFDIALOGUE	FISHER	HELLASWAG	PDTB
training	40154	52665	16757	4351	14738	1743
validation	4989	1571	3592	930	1600	374
test	5054	1571	3594	934	1643	374

Table 1: Number of documents in the training, validation, and test portions of our six corpora.

architecture composed of two components: First component is a biLSTM sentence encoder that uses Elmo (Peters et al., 2018) embeddings as input; pairs of adjacent sentence vectors are then passed through a bilinear layer to obtain a combined vector representation of these. Second component is a lightweight CNN with average pooling (Wu et al., 2019) that takes as input the sentence embeddings generated by the sentence encoder, and outputs a vector summarizing the entire document. These two vectors are then concatenated and passed through a linear layer to output a local coherence score. The global coherence score for a document is calculated by summing up the local scores.

5 Experimental setup

5.1 Corpora

We carry out extensive evaluation on corpora from a number of different domains, namely:

***Narration*:** VISUAL STORYTELLING (VIZSIS) (Huang et al., 2016) and ROCSTORIES (Mostafazadeh et al., 2016) both contain crowdsourced short stories containing ∼5 sentences, where participants are prompted with a visual/verbal cue, and are asked to generate a coherent story containing the relevant entities and events.

***Dialogue*:** FISHER (Cieri et al., 2004) contains phone conversations between two speakers whereas SELFDIALOGUE (Fainberg et al., 2018) contains fictional conversations imagined by a speaker. We choose these two corpora given that they prompt the speakers for conversations around well-defined topics, ranging from everyday subjects, such as food, family, and friends, to more specific issues/entities, such as airport security, Lady Gaga, etc.

***Instructional text*:** HELLASWAG (Zellers et al., 2019) is a collection of instructional short texts from WikiHow, combined with short temporal descriptions of everyday human activities from ActivityNet Captions (Krishna et al., 2017). The dataset was originally compiled for the task of common-sense reasoning, and contains one correct ending (last sentence) and three automatically-generated incorrect endings (distractors) per document.

***News*:** We also conduct experiments on the Penn Discourse Treebank (PDTB) portion (Prasad et al., 2008) of the Wall Street Journal corpus, as commonly used for the evaluation of coherence models.

The VIZSIS, HELLASWAG, and ROCSTORIES[6] corpora come already partitioned into training, validation and test sets. For the remaining corpora we adopt a 70/15/15 training/validation/test set split. For ROCSTORIES in story cloze task, we need access to the correct and incorrect endings of each story for evaluation. Since this information is only available for the validation portion, we use this as our final test set. In all our experiments, we use the validation set to choose the best epoch for the models, and report the results on the test set for that epoch. The only exception is ROCSTORIES story cloze, where we use the validation set both for selecting the best epoch and for reporting accuracy. The number of documents in each set for every corpora is provided in Table 1.

For all corpora and tasks we filter out documents with less than 2 sentences, except for topic switching where we set the minimum document length to 4 for LDA to learn robust topic distributions. We tune the parameters of the LDA topic modeling independently for each corpora. These parameters include the number of topics, and the percentage of top-ranked words based on TF-IDF scoring to be used as topic vocabulary. Final values of these two parameters are reported in Appendix A, along with the total number of training/validation/test pairs.

[6]We use the winter 2017 training set and the StoryCloze Test Winter 2018 validation set available upon request.

	VIZSIS	ROCSTORIES	SELFDIALOGUE	FISHER	HELLASWAG	PDTB
ParSeq	75.35	77.84	88.40	93.48	**86.67**	91.00
CohLSTM	82.25	89.55	90.79	**99.20**	69.38	61.96
UNC	**88.42**	**94.80**	**97.21**	n/a	83.92	**92.85**

Table 2: Accuracy of models on order discrimination, excluding SentAvg as it does not capture order.

	same-full	*same-half*	*different-full*	*different-half*
SentAvg	74.62	67.93	79.74	68.06
ParSeq	70.7	61.59	63.71	57.68
CohLSTM	84.99	78.76	79.04	74.11
UNC	**96.46**	**96.77**	**95.11**	**96.89**

Table 3: Accuracy of all models for the connective substitution task on the PDTB test set.

5.2 Model implementations and evaluation metric

For the models in our study (explained in Section 4), we use the codes provided by the authors,[7] with minimal changes required to enable working with the different datasets. We do not do any hyperparameter tuning and use the reported hyperparameters in the corresponding papers. Following previous work, for every task we create up to 20 incoherent texts per coherent document. In all experiments, we use accuracy to assess the performance of the models. Every sample in our data consists of a pair of {*coherent, incoherent*} documents, and thus accuracy for a model is measured as the number of times that the model ranks a coherent document in a pair higher than the incoherent one, divided by the total number of test pairs.

6 Results

6.1 Order discrimination

We start by presenting results on the standard order discrimination task to better understand the connection between the performance on this vs. the tasks we designed. Results for the task of order discrimination are given in Table 2.[8] Our results generally confirm that the UNC model of Moon et al. (2019) is indeed SOTA on this particular task: it outperforms the other models on five out of the six corpora, with a large margin on three of them (VIZSIS, ROCSTORIES, and SELFDIALOGUE). Both UNC and CohLSTM show a notable drop in performance on HELLASWAG, while the accuracy of CohLSTM is above 99% on FISHER. This is indeed an interesting observation: HELLASWAG contains short stories, with more than 90% of them being 2–3 sentences long, whereas FISHER has very long documents, with an average of more than 150 sentences per document. As noted by Elsner and Charniak (2011), order discrimination is easier for longer documents, especially for models that have been optimized for this task. CohLSTM also shows a significant drop in performance on PDTB, while ParSeq performs poorly on the two narrative datasets (VIZSIS and ROCSTORIES).

6.2 Connective substitution

Results for the task of connective substitution are shown in Table 3; because for this task we need to identify connectives, we only report results on the PDTB corpus that contains such annotations. Interestingly, the UNC model outperforms all other models by a large margin: UNC specifically models global coherence patterns using a convolution-pooling mechanism, which we believe enables the model to capture correct vs. incorrect use of connectives across sentences. As expected, it is generally easier to identify incoherent

[7] For SentAvg and ParSeq, we follow `https://github.com/aylai/GCDC-corpus` as suggested by the authors. For CohLSTM, we use the online repo provided at `https://github.com/MMesgar/neural_coherence_model`, and for UNC we use the provided code at `https://github.com/taasnim/unified-coherence-model`.

[8] Due to its complexity, UNC does not run on very long documents with a length of greater than 100; the original model filters out such documents. In both training and test portions of FISHER, more than 95% of documents are thus filtered out, and as such we cannot report reliable performance numbers for UNC on this data set.

	VIZSIS	ROCSTORIES	SELFDIALOGUE	FISHER	PDTB
controlled mixing (*random mixing*)					
SentAvg	71.87 (63.28)	74.17 (54.08)	80.54 (54.20)	**83.02** (50.96)	48.56 (49.33)
ParSeq	82.85 (85.37)	91.99 (90.73)	**81.46** (**87.44**)	72.53 (48.66)	70.33 (**65.56**)
CohLSTM	64.81 (52.23)	67.85 (59.09)	68.41 (53.74)	82.55 (**53.05**)	52.33 (47.89)
UNC	**92.10** (**88.40**)	**94.62** (**92.20**)	71.74 (83.85)	n/a	**70.89** (61.73)

Table 4: Accuracy for both topic switching conditions, with *random mixing* results reported in parentheses.

documents that arise from replacing connectives in all sentences (*full* conditions; columns 2 & 4 from the left); this is the case for all models except UNC that performs well across all conditions. No consistent patterns are observed between the *same* and *different* conditions.

We expected the incoherent documents arising from connective substitution to be harder to discriminate compared to the shuffled documents. Comparing the results on the connective substitution task (Table 3) with those on order discrimination on PDTB (Table 2; right-most column), we can see that while this is the case for ParSeq (with substantially lower performance on connective substitution), it does not hold true for the other two models: Whereas UNC performs equally well on both tasks, CohLSTM performs better on the connective substitution task. Some initial analysis of our data generation process for this task shows biases in the substitution of high-frequency connectives (e.g., *and*, *but*, *also*) — that is, not only there are a few very high-frequency connectives, but also they tend to be mainly substituted with only a few other high-frequency connectives. For instance, more than 85% of occurrences of *and* are substituted by three high-frequency connectives, namely *but*, *or*, and *as*. Further analysis is required to understand whether any component of CohLSTM or UNC makes them susceptible to using this bias to their advantage, hence performing well on this task without actually learning about coherence.

6.3 Sentence cloze

Results for the task of sentence cloze can be seen in Table 5a where we report results on PDTB. On this task, CohLSTM outperforms the other models, including UNC, with a notable margin. Recall that in sentence cloze we specifically manipulate lexical cohesion, and CohLSTM is a model designed particularly to capture this aspect of coherence through "encod[ing] the perceived coherence of a text by a vector, which represents patterns of changes in salient information that relates adjacent sentences" (Mesgar and Strube, 2018). Indeed, CohLSTM performs much better on sentence cloze compared to order discrimination on the same corpus (cf. right-most column of Table 2). Interestingly, both SentAvg and ParSeq perform relatively poorly on this task, most likely because this information is lost either due to averaging across all word vectors (SentAvg) or due to the complex multi-layer recurrent architecture (ParSeq).

6.4 Topic switching

Results for the task of topic mixing are given in Table 4; we exclude HELLASWAG from these experiments because its documents tend to be very short. For all models and across all corpora, performances in the controlled mixing condition are generally higher or comparable to those for random mixing. This is in line with our expectation that the random mixing would be the harder of the two conditions: it is possible for randomly-mixed documents to have similar topics since we do not control for their degree of topic overlap. For this condition, the two best-performing models are UNC and ParSeq (with the exception of FISHER), both of them showing performance drops on corpora with long documents, namely FISHER (with an average document length of ~ 150) and PDTB (with an average of ~ 30 sentences per document). This is an interesting, yet expected, observation: when we mix two halves from long documents, the resulting document is still somewhat coherent as the two halves would have a big impact on the overall coherence of the combined document. In both conditions, CohLSTM generally performs poorly on this task, although it has been designed to capture *local* topic transitions.

	full substitution	*half substitution*
SentAvg	69.47	59.63
ParSeq	48.48	59.22
CohLSTM	**98.19**	95.88
UNC	94.41	**96.00**

(a) Accuracy for the sentence cloze task on PDTB test set.

	HELLASWAG	ROCSTORIES	
		original	*random*
SentAvg	**56.32**	50.86	75.18
ParSeq	55.97	50.48	81.41
CohLSTM	51.31	49.97	**92.62**
UNC	34.03	**73.87**	92.22

(b) Accuracy for story cloze task.

Table 5: Accuracy for the two cloze tasks.

6.5 Story cloze

Results for the task of story cloze are shown in Table 5b. Here, we report results for the original test set of HELLASWAG as well as the original validation set of ROCSTORIES that contain a correct ending and one or more incorrect endings.[9] In addition, we also include results on a modified version of the ROCSTORIES validation set where we generate a *random* incorrect ending instead of the originally-provided incorrect ending. The reason is that the training portion of this data set does not include human-generated incorrect endings, and as such we do generate those randomly.[10] We thus provide results on a similarly generated set for comparison. Of course we expect the *random* endings to be much easier to identify compared to the *original* human-written endings, which is confirmed by the much higher accuracies of all models for this condition compared to the *original* condition.

On the original story cloze task, we can see that all models perform much worse than other tasks (order discrimination, connective substitution, sentence cloze and topic switching). This is expected as story cloze is designed to require some degree of common-sense reasoning and inference, in addition to understanding coherence. Still, the UNC model of Moon et al. (2019) has the best performance on ROCSTORIES ($\sim 74\%$ accuracy on the *original* task). The rather poor performance of UNC on HELLASWAG is possibly due to document length, and the reliance of UNC on contextual information. Note that HELLASWAG has been designed to be particularly difficult for deep learning models by providing little contextual information (Zellers et al., 2019). Although we can see that a simple model such as SentAvg can still do reasonably well even without any common-sense reasoning (note the baseline performance on HELLASWAG is 25% since there are three incorrect endings per correct ending for each test story).

7 Discussion and conclusions

Our results point to a first important conclusion: the task of order discrimination paints only a partial picture of what is SOTA in modelling coherence, as also noted by other researchers (Lai and Tetreault (2018) *inter alia*). This conclusion is evident from the high variability in performance of models across tasks and corpora. Particularly, we observe that some models exhibit clear advantage over others on certain tasks and/or corpora, e.g., CohLSTM does particularly well on the sentence cloze task which manipulates the degree of lexical cohesion in a document. In other words, *different models seem to fit some tasks and corpora better than the others*. In addition, *certain properties of the corpora affect different models differently*. We observe that document length is of particular importance for models that represent the overall coherence of an entire document either via convolution (UNC), or LSTM (ParSeq), as attested by their results on *document-level* tasks, namely topic switching and story cloze. For topic mixing, this is observed in the two corpora with long documents (FISHER and PDTB), and for story cloze, by the poor performance on HELLASWAG.

Finally, *all models still struggle with common-sense reasoning*, especially when there is little context available. We believe this kind of common-sense reasoning over short stories is an integral aspect of overall coherence, and as such we see this relevant to coherence modeling. Results comparing randomly-generated incorrect endings vs. human generated ones strengthen our initial claim: the fact that random

[9] As noted previously, we report these results on the validation set of ROCSTORIES since their test set does not come with the annotation of correct/incorrect endings.

[10] We do this by substituting the last sentence in the positive document with a random *final* sentence from another document.

alternatives/distractors can be easily spotted by existing SOTA models does not necessarily mean that these models are capturing true aspects of coherence that human readers are sensitive to.

All in all, our findings have important implications, both for the design of coherence models, and for their proper evaluation. Specifically, the two should not be seen as independent goals, but rather they should inform each other: by designing appropriate and diverse evaluation tasks, we can design multi-faceted models that capture all aspects of coherence. At the same time better models require better evaluation techniques that clearly showcase their strengths and shortcomings.

Future work should aim at extending this line of work to other domains. For instance, these tasks could be easily applied to measure the coherence of texts produced by L2 learners or on machine translation output. Furthermore, the tasks we propose could be further refined. For instance, in the Sentence Cloze task we pooled hypernyms, hyponyms and antonyms together to generate distractors but models might be more sensitive to one of these classes; similarly, the topic switching did not look at the similarity between topics which could also affect model performance. Finally, one could also experiment with even lower levels of distortion than just half the sentences in the document.

Acknowledgments

We kindly thank the three anonymous reviewers for their useful comments. Research was conducted at Samsung AI Centre Toronto and funded by Samsung Research, Samsung Electronics Co., Ltd.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Annual Meeting of the Association for Computational Linguistics*, pages 789–797.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: a resource for the next generations of speech-to-text. In *the International Conference on Language Resources and Evaluation*, volume 4, pages 69–71.

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? finding the optimal feature set for general readability prediction and its application to english and dutch. *Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Tony Deyes. 1984. Towards an authentic discourse cloze. *Applied Linguistics*, 5(2).

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129, Portland, Oregon, USA, June. Association for Computational Linguistics.

Joachim Fainberg, Ben Krause, Mihai Dobre, Marco Damonte, Emmanuel Kahembwe, Daniel Duma, Bonnie Webber, and Federico Fancellu. 2018. Talking to myself: self-dialogues as data for conversational agents. *arXiv preprint arXiv:1809.06641*.

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *International Conference on Computational Linguistics*.

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

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Yacine Jernite, Samuel R. Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *the IEEE international conference on computer vision*, pages 706–715.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation, and methods. In *the Annual SIGdial Meeting on Discourse and Dialogue*.

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

Jiwei Li and Dan Jurafsky. 2017. Neural net models for open-domain discourse coherence. In *Empirical Methods in Natural Language Processing*.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Annual Meeting of the Association for Computational Linguistics*.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. *arXiv preprint arXiv:1611.02654*.

Annie Louis and Ani Nenkova. 2012. A coherence model based on syntactic patterns. In *Empirical Methods in Natural Language Processing*.

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Empirical Methods in Natural Language Processing*.

Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. 2019. Unsupervised learning of discourse-aware text representation for essay scoring. In *ACL Student Workshop*.

Tasnim Mohiuddin, Shafiq Joty, and Dat Tien Nguyen. 2018. Coherence modeling of ascynchronous conversations: a neural entity grid approach. In *Annual Meeting of the Association for Computational Linguistics*.

Han Cheol Moon, Tasnim Mohiuddin, Shafiq Joty, and Xu Chu. 2019. A unified neural coherence model. In *Empirical Methods in Natural Language Processing*, pages 2262–72.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48.

Nasrin Mostafazadeh, Nathanae Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Dat Tien Nguyen and Shafiq Joty. 2017. A neural local coherence model. In *Annual Meeting of the Association for Computational Linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Empirical Methods in Natural Language Processing*.

R. Prasad, A. Lee, N. Dinesh, E. Miltsakaki, G. Campion, A. Joshi, and B. Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *International Conference on Computational Linguistics*, pages 950–961.

Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2018. Automatic distractor generation for multiple-choice english vocabulary questions. *Research and Practice in Technology Enhanced Learning*, 13(15).

Wilson L. Taylor. 1953. Cloze procedure: a new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *American Association for Artificial Intelligence*.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *the Annual Meeting of the Association for Computational Linguistics*.

Muyu Zhang, Vanessa Wei Feng, Bing Qin, Graeme Hirst, Ting Liu, and Jingwen Huang. 2015. Encoding world knowledge in the evaluation of local coherence. In *North American Chapter of the Annual Meeting of the Association for Computational Linguistics*.

A Topic switching - parameters and data statistics

	VIZSIS	ROCSTORIES	SELFDIALOGUE	FISHER	PDTB
# topics	30	30	30	30	20
top-%	50	50	50	15	20
# training pairs	505,800 (12.6)	665,860 (12.6)	187,724 (11.2)	55,080 (12.7)	18,420 (12)
# validation pairs	62,480 (12.6)	19,680 (12.5)	13,726 (3.8)	2,311 (2.5)	3,804 (12)
# test pairs	64,340 (12.7)	20,100 (12.8)	14,600 (4.1)	6,992 (7.5)	3,600 (12)

Table 6: Parameter values of the LDA model, and number of pairs in each training/validation/test split. The numbers in parentheses show the average number of negative samples generated per positive.