Rethinking Coherence Modeling: Synthetic vs. Downstream Tasks

Tasnim Mohiuddin[*¶], **Prathyusha Jwalapuram**[*¶], **Xiang Lin**[*¶], and **Shafiq Joty** [*¶†]

[¶]Nanyang Technological University, Singapore
[†]Salesforce Research Asia, Singapore
{mohi0004, jwal0001, linx0057, srjoty}@ntu.edu.sg

Abstract

Although coherence modeling has come a long way in developing novel models, their evaluation on downstream applications for which they are purportedly developed has largely been neglected. With the advancements made by neural approaches in applications such as machine translation (MT), summarization and dialog systems, the need for coherence evaluation of these tasks is now more crucial than ever. However, coherence models are typically evaluated only on synthetic tasks, which may not be representative of their performance in downstream applications. To investigate how representative the synthetic tasks are of downstream use cases, we conduct experiments on benchmarking well-known traditional and neural coherence models on synthetic sentence ordering tasks, and contrast this with their performance on three downstream applications: coherence evaluation for MT and summarization, and next utterance prediction in retrieval-based dialog. Our results demonstrate a weak correlation between the model performances in the synthetic tasks and the downstream applications, motivating alternate training and evaluation methods for coherence models.[1]

1 Introduction and Related Work

Coherence is an important aspect of discourse that distinguishes a well-written text from a poorly-written one that is difficult to comprehend (Halliday and Hasan, 1976). Computational models that can assess coherence have applications in text generation and ranking, such as summarization, machine translation, essay scoring and dialog systems.

Researchers have proposed a number of formal theories of discourse coherence, which have in-

spired the development of many coherence models – both traditional and neural ones. Inspired by the Centering Theory (Grosz et al., 1995), the *entity* based local models (Barzilay and Lapata, 2008; Elsner and Charniak, 2011b) formulate coherence in terms of syntactic roles (*e.g.,* subject, object) of entities in nearby sentences. Another branch of models (Pitler and Nenkova, 2008; Lin et al., 2011; Feng et al., 2014) use *coherence relations* between adjacent sentences to model local coherence, inspired by the discourse structure theories of Mann and Thompson (1988) and Webber (2004). Other traditional methods include *word co-occurrence* based local models (Soricut and Marcu, 2006), *topic* based global models (Barzilay and Lee, 2004; Elsner et al., 2007), and *syntax* based local and global models (Louis and Nenkova, 2012).

Despite continuous research efforts in developing novel coherence models, their usefulness in downstream applications has largely been ignored. They have been evaluated in mainly two ways. The most common approach has been to evaluate them on *synthetic* discrimination tasks that involve identifying the right order of the sentences at the local and global levels (Barzilay and Lapata, 2008; Elsner and Charniak, 2011b; Moon et al., 2019). The other (rather infrequent) way has been to assess the impact of coherence score as an additional feature in downstream tasks like *readability assessment* and *essay scoring* (Barzilay and Lapata, 2008; Mesgar and Strube, 2018). But since the concept of coherence goes beyond these constrained tasks and domains, so should the models.

Given the recent advances in neural NLP methods, with claims of reaching human parity in machine translation (Hassan et al., 2018), fluency in summarization (Liu et al., 2017; Celikyilmaz et al., 2018), or context-consistent response generation (Zhang et al., 2020; Hosseini-Asl et al., 2020), coherence modeling of machine-generated texts, par-

---

*Equal contribution

[1]Code and data used for evaluation available at https://ntunlpsg.github.io/project/coherence/coh-eval/

ticularly at a document-level, is now more crucial than ever (Läubli et al., 2018; Sharma et al., 2019). Traditional task-specific evaluation methods (*e.g.,* BLEU, ROUGE) may not be an accurate reflection of their real-world performance in terms of readability (Paulus et al., 2017; Reiter, 2018). However, it is unclear if existing coherence models are capable of this task, since their performance on downstream applications is rarely studied, even though that is one of the main motivations for their development.

Our main goal in this work is to assess the performance of the existing coherence models not only on standard, challenging synthetic tasks like global and local discrimination, but more importantly on real downstream text generation problems. Specifically, we investigate the performance of coherence models in three different settings:

- Traditional synthetic tasks involving discrimination of real documents from their permutations.

- Coherence evaluation for machine translations and system-generated extractive and abstractive summaries, which are more representative of real-world use cases for coherence models.

- Next utterance ranking for dialogs, which is a downstream application similar to the synthetic task of insertion, but uses conversational data from DSTC 8 (Kim et al., 2019).

We show through experiments that there is only a slight correlation between model performances on synthetic tasks and the real-world use cases. Although models perform strongly in the synthetic tasks, they show poor performance and low correlations with human judgments on distinguishing coherent machine translations and system-generated summaries from incoherent ones. They also fail to perform well on the next utterance ranking task, which is similar to the synthetic task of insertion (Elsner and Charniak, 2011b), even if re-trained with task-specific data.

However, we show that re-training the coherence models with task-specific data for machine translation evaluation leads to improved results and agreements with human judgments. This leads us to conclude that there is a possible mismatch in the task setting that is used to train coherence models. Models trained on traditional synthetic tasks do not seem to be learning features that are useful for downstream applications. We hope that our results will motivate the broadening of the standard

of coherence model evaluations to include more downstream tasks, and also motivate the redesigning of the training paradigm for coherence models.

2 Coherence Models

Advancements in deep learning have inspired researchers to neuralize many of the traditional models. Li and Hovy (2014) model syntax and inter-sentence relations using a recurrent sentence encoder followed by a fully-connected layer. In a follow-up work, Li and Jurafsky (2017) use generative models to incorporate global topic information with an encoder-decoder architecture. Mohiuddin et al. (2018) propose a neural entity grid model using convolutions over distributed representations of entity transitions. Mesgar and Strube (2018) model change patterns of salient semantic information between sentences. Xu et al. (2019) propose a local discriminative model that retains the advantages of generative models and uses a smaller negative sampling space that can learn against incorrect orderings. Moon et al. (2019) propose a unified model that incorporates sentence syntax, inter-sentence coherence relations, and global topic structures in a single Siamese framework.

We benchmark the performance of five representative coherence models on the tasks discussed above. Our selected models comprise of both traditional and neural models. Moreover, two models are currently the state-of-the-art at the time of submission (Transferable and Unified Neural Model).

Entity Grid (EGRID). Barzilay and Lapata (2005, 2008) introduced the popular entity-based model for representing and assessing text coherence motivated by the Centering Theory (Grosz et al., 1995). This model represents a text with a two-dimensional array called an *entity grid*, that captures transitions of discourse entities across sentences. These local entity transitions are used as deciding patterns for text coherence; a local entity transition of length k is a sequence of $\{S,O,X,-\}^k$ representing grammatical roles (*Subject*, *Object*, *Other*, and *Absent*, respectively) played by an entity in k consecutive sentences. The *salience* of the entities, quantified by the occurrence frequency, is also incorporated to identify transitions of important entities. Elsner and Charniak (2011b) improve the basic entity grid by including non-head nouns as entities (with the grammatical role X). Instead of using a coreference resolver, they match the nouns

3529

to detect coreferent entities. In our work, we consider this version of the entity grid model.

Neural Entity Grid (NEURALEGRID**).** A neural version of the entity grid model was proposed by Nguyen and Joty (2017). The grammatical roles in the grid are converted into their distributed representations, and the entity transitions are modeled in the distributed space by performing convolutions over it. The final coherence scores are computed from convolved features that have gone through a spatial max-pooling operation. A global, document-level pairwise loss is used to train the model.

Lexicalized Neural Entity Grid. Mohiuddin et al. (2018) propose an improvement of the neural entity grid (LEXNEUEGRID) by *lexicalizing* the entity transitions using off-the-shelf word embeddings to achieve better generalization.

Transferable Neural Model (TRANSMODEL**).** In order to generalize the coherence model across domains, Xu et al. (2019) propose a transferable neural model that considers coherence at a local level, taking only adjoining sentences as input. Coupled with pre-training of the sentence encoders in a generative fashion, their model demonstrates significant improvements in performance, despite being a local coherence model.

Unified Neural Model (UNIFIEDMODEL**).** Moon et al. (2019) propose a unified model that captures syntax (as a proxy of intention), discourse relations, entity attention and global topic structures. The syntax is captured by incorporating an explicit language model loss. A bi-linear layer is used to capture the inter-sentential discourse relations, while light-weight convolution is used to capture the attention and topic structures.

3 Evaluation Tasks and Experiments

In this section, we present the performance of the coherence models on standard synthetic tasks (*i.e.,* Global/Local Discrimination), followed by the experiments where we apply the coherence models trained on the global discrimination task to three downstream tasks (*i.e.,*, abstractive summarization, extractive summarization, and machine translation). We then present the results of the coherence models re-trained on the next utterance ranking task.

For each of the coherence models, we conducted experiments with publicly available codes from the respective authors. The three recent methods

	Sections	# Doc.	# Pairs
Train	00-13	1,378	26,422
Test	14-24	1,053	20,411

Table 1: Statistics of the WSJ news dataset used for the **Global discrimination** task.

use word embeddings: LEXNEUEGRID, TRANSMODEL and UNIFIEDMODEL use Word2vec (Mikolov et al., 2013), average GloVe (Pennington et al., 2014), and ELMo (Peters et al., 2018) embeddings respectively. We use the default settings and hyperparameters suggested by the authors.

3.1 Synthetic Tasks

Traditionally coherence models have been evaluated mostly on synthetic tasks. For comparison with previous work, we use two representative synthetic tasks to compare the coherence models.

3.1.1 Global Discrimination.

Introduced by Barzilay and Lapata (2008), in this task coherence models are asked to distinguish an original (coherent) document from its incoherent renderings generated by random permutations of its sentences. We follow the same experimental setting of the Wall Street Journal (WSJ) news dataset as used in previous studies (Elsner and Charniak, 2011b; Moon et al., 2019; Xu et al., 2019). Similar to them, we use 20 random permutations of each document for both training and testing. Additionally, we evaluate on *inverse discrimination* (Mohiuddin et al., 2018), where the sentence order is reversed to create the incoherent version.

Setup. We follow the same experimental settings of the WSJ news dataset as used in previous works (Xu et al., 2019; Mohiuddin et al., 2018; Elsner and Charniak, 2011b; Feng et al., 2014). We use 20 random permutations of each document for both training and testing, excluding the permutations that match the original one. Table 1 summarizes the data sets used in the global discrimination task. We randomly select 10% of the training set for development purposes.

Results. Table 2 presents the results in terms of accuracy on the two *global discrimination* tasks – the *standard* and the *inverse* order discrimination. We see that UNIFIEDMODEL achieves the highest accuracy on the standard order discrimination task and TRANSMODEL performs the best on the Inverse order discrimination task. The other three

Model	Emb.	Standard	Inverse
EGRID	–	81.60	75.78
NEURALEGRID	–	84.36	83.94
LEXNEUEGRID	word2vec	88.51	88.13
TRANSMODEL	Avg. Glove	91.77	**99.62**
UNIFIEDMODEL	ELMo	**93.19**	96.78

Table 2: Results: Accuracies of the coherence models in the **Global Discrimination** task.

	Sections	# Doc.	# Pairs			
			$\mathcal{D}_{w=1}$	$\mathcal{D}_{w=2}$	$\mathcal{D}_{w=3}$	$\mathcal{D}_{w=1,2,3}$
Train	00-13	748	7,890	12,280	12,440	32,610
Test	14-24	618	6,568	9,936	9,906	26,410

Table 3: Statistics on the WSJ news dataset used for the **Local discrimination** task. The w denotes the number of permuted local windows in a document.

models use entity grids, hence they may lose the sentence-level syntactic and semantic information.

3.1.2 Local Discrimination.

Local discrimination was proposed by Moon et al. (2019). In this task, two documents differ only in a local context (windows of 3 sentences). In this case, the models need to be sensitive to local changes. We use the same WSJ dataset as used by Moon et al. (2019).

Setup. We use the same WSJ articles used in the global discrimination task (Table 1) to create our local discrimination datasets. We use the code released by Moon et al. (2019) to generate these datasets.[2] Sentences within a local **window of size 3** are re-ordered to form a locally incoherent text. Only articles with **more than 10 sentences** are included in the dataset. Table 3 summarizes the datasets. We randomly select 10% of the training set for development purposes.

Following Moon et al. (2019), we create four datasets for our local discrimination task: $\mathcal{D}_{w=1}$, $\mathcal{D}_{w=2}$, $\mathcal{D}_{w=3}$ and $\mathcal{D}_{w=1,2,3}$. $\mathcal{D}_{w=1}$ contains the documents where only one randomly selected window is permuted, $\mathcal{D}_{w=2}$ contains the documents where two randomly selected windows are permuted; $\mathcal{D}_{w=3}$ is similarly created for 3 windows. $\mathcal{D}_{w=1,2,3}$ denotes the concatenated datasets.

Results. From Table 4, we see that the UNIFIED-MODEL achieves the highest accuracy on all four datasets. A possible reason could be the loss function it uses to train the model. Unlike other models, UNIFIEDMODEL uses an adaptive pairwise ranking

---
[2] https://github.com/taasnim/unified-coherence-model

Model	$\mathcal{D}_{w=1,2,3}$	$\mathcal{D}_{w=1}$	$\mathcal{D}_{w=2}$	$\mathcal{D}_{w=3}$
EGRID	59.78	53.89	60.43	63.04
NEURALEGRID	57.49	56.74	57.11	60.0
LEXNEUEGRID	56.65	58.21	58.95	58.42
TRANSMODEL	66.87	66.25	67.95	65.52
UNIFIEDMODEL	**77.07**	**67.29**	**76.12**	**81.23**

Table 4: Results: Accuracies of the models in the **Local Discrimination task**.

loss which does not penalize the locally coherent sentences. In the local discrimination task, the difference between positive and negative examples is small; they differ only in 1-3 windows, while the other parts are locally coherent. UNIFIEDMODEL's loss function can model this better.

3.2 Coherence Evaluation Tasks

We evaluate the coherence models trained on the global discrimination task on two downstream tasks: machine translation (MT) and summarization coherence evaluation. Note that both the MT and summarization data are from the same domain (news) as the original WSJ training data.

3.2.1 Machine Translation Evaluation

The outputs of neural machine translation (NMT) systems have been shown to be more fluent than their phrase-based predecessors (Castilho et al., 2017). However, recent studies have shown that there is a statistically strong preference for human translations in terms of both adequacy and fluency at a document level (Läubli et al., 2018; Popel et al., 2020).

Smith et al. (2016) evaluated traditional (non-neural) coherence models to see if they can distinguish a reference from a system translated document, and reported very low accuracy. However, the situation has changed with the advancements of neural models; today's coherence models are claimed to be much more accurate.

Our goal therefore is to evaluate the coherence models on how well they can judge the coherence of MT outputs at the document level. To do this, we use the system translations released by the annual Workshop (now Conference) on Machine Translation (WMT) through the years 2017 and 2018. At a document level, reference (human) translations have been shown to be more coherent than MT outputs (Smith et al., 2015, 2016; Läubli et al., 2018). Therefore, we evaluate the performance of the coherence models based on their accuracy of scoring

Candidate A:
"A Generation Is Protesting" in Ethiopia, Long a U.S. Ally Violent Protests in Ethiopia Demonstrators demanding political change in Ethiopia have been met with violent resistance by the government. Witnesses say that scores of protesters have been fatally shot during clashes with police.

`1 ⌄`

Candidate B:
"A Nesil Protestode" in Etiyopia, the United States, for a long time. Violent Protests in Ethiopia Demonstrators demanding political change in Ethiopia faced severe resistance by the government. Several protesters have taken deadly weapons coup in the conflict with police, Sahitler says.

`1 ⌄`

Candidate C:
"In a generation protest" in Ethiopia, an ally of the United States for a long time Protests involving violence in Ethiopia Demonstrators who demanded political change in Ethiopia faced the government's fierce resistance. Witnesses say a large number of protesters received deadly weapons in the shootout with the police.

Figure 1: User study interface for coherence ranking.

Model	Acc. (%)	AC1 Agr.
EGRID	51.75	0.80
NEURALEGRID	54.75	0.77
LEXNEUEGRID	49.34	0.76
TRANSMODEL	48.67	0.77
UNIFIEDMODEL	43.36	0.78

Table 5: **Machine Translation** setting results on WMT2017-2018 data. **Acc**uracies: % of times reference scored higher and **AC1 agr**eements for system translation rankings between annotators and models.

the reference (document) higher than the system translation (document).

We also obtain rankings given by humans in a user study. Fig. 1 shows the layout of the study, where participants were shown four sentences from three candidate translations of the same source text and asked to rank them against each other. One of the given translations is the reference, used as a control, and to validate our assumption that the reference is more coherent than the system translations. 3 participants annotated 100 such samples.

Participants chose the reference as more coherent with an agreement of **0.84**, confirming our assumption.[3] We evaluate the system translations by producing a ranking between the different translations of the same source text. To do this, we first obtain scores from the coherence models for the reference and each of the corresponding system translations. Then, we *normalize* the scores of the system translations by subtracting them from score of the reference. These normalized coherence scores are used to rank the system translations, which are then used to calculate agreements.

Setup. We use the reference and the system translations provided by WMT2017-2018 as our test data, under the assumption that the reference translations are more coherent than the system translations. This results in a testset of 20,680 reference-system translation document-pairs.

Results. We report the accuracy of the coherence models trained on the global discrimination task in distinguishing the more coherent reference text from the less coherent system translations in Table 5. We can see that most models perform

worse than a random baseline of 50%, showing that their training on the global discrimination task is not helpful in detecting coherence quality in MT text. The difference in performance is particularly glaring for the TRANSMODEL and the UNIFIED-MODEL, both of which have over 90% accuracy on the global discrimination tasks, but only manage 48.67% and 43.36% on this task respectively.

We also report the agreement with human rankings on the study data in Table 5. Overall, only EGRID has good agreement with human rankings, with all other models doing similarly poorly.[4]

3.2.2 Abstractive Summarization

Generating coherent summaries has always been a goal in summarization (Nenkova and McKeown, 2011). The widely used automatic evaluation metric ROUGE (Lin, 2004) measures the n-gram overlap between the generated summaries and the reference summaries at a sentence level, and thus is not sufficient for measuring coherence. Kryściński et al. (2019) also recently found almost negligible correlation between ROUGE scores and human judgments on summary coherence, especially for abstractive summaries generated by recent neural summarization models. We therefore propose to evaluate the coherence of summaries using different coherence models and measure their effectiveness on this task.

For abstractive summarization, we use summaries from popular neural abstractive summarization systems for CNN/DM dataset (Hermann et al., 2015; Nallapati et al., 2016). Since abstractive systems vary in their architectures and loss functions, they may produce very different summaries. We run a human study to validate the rankings given by the coherence models.

---

[3]Traditional correlation measures such as Cohen's Kappa are not robust to skewed distributions of annotations, which was an issue here since the annotators were always more likely to choose the reference as better. Thus, we report the more appropriate Gwet's AC1/gamma coefficient (Gwet, 2008), which controls for this.

[4]Note that the study data is different from the test data, so the accuracies and agreements may not correlate.

Models	Abs. Agr.	Ext. Agr.
EGRID	**0.71**	0.52
NEURALEGRID	0.68	**0.70**
LEXNEUEGRID	**0.71**	0.57
TRANSMODEL	0.55	0.38
UNIFIEDMODEL	0.68	0.35

Table 6: **Abs**tractive **Agr**eement and **Ext**ractive **Agr**eement shows the AC1 agreements for the pairwise ranking of the generated abstractive summaries and extractive summaries between two annotators and the models, respectively.

Setup. We use the CNN/DM (Hermann et al., 2015; Nallapati et al., 2016) for this task. We collect the reference summaries from the CNN/DM testset as well as the summaries generated by the following four representative abstractive summarization systems: (*a*) Pointer-Generator (PG) (See et al., 2017), (*b*) BERTSUMEXTABS (BSEA) (Liu and Lapata, 2019), (*c*) UniLM (Dong et al., 2019), and (*d*) SENECA (Sharma et al., 2019).

As discussed, we directly use the coherence models trained on the WSJ dataset for the global discrimination task. The coherence models predict the scores for each system-generated summary in the testset. The scores produced by the models are then used to rank the system-generated summaries of the same original article.

We conducted a user study to validate the effectiveness of the rankings produced by the coherence models. We randomly sampled 10 sets of summaries from the dataset with each set containing four generated summaries of the same article, thus resulting in $\binom{4}{2} \times 10 = 60$ pairs of system summaries. Two annotators were asked to rank each pair of the summaries in terms of coherence; see Appendix for the human study interface.

Results. For the user study, the agreement between the two annotators was **0.78**, which indicates fairly reliable data. After we obtain the rankings based on the coherence scores produced by the models, we compute the agreements between the systems and the two annotators. From the results in Table 6, we see that EGRID and LEXNEUEGRID show the highest agreement with human judgements. However, despite strong performance in synthetic tasks, models like UNIFIEDMODEL and TRANSMODEL are unable to convert the high accuracy into high human agreement, which demonstrates the inefficiency of current synthetic tasks.

3.2.3 Extractive Summarization

For evaluating the coherence of extractive summaries, we use the dataset prepared by Barzilay and Lapata (2008) for their coherence model evaluation. The dataset comes with human ratings of the summaries from the Document Understanding Conference (DUC), 2003.

Setup. The dataset from Barzilay and Lapata (2008) provides 16 sets of summaries where each set corresponds to a multi-document cluster and contains summaries generated by 5 systems and 1 human. The human ratings for these summaries based on coherence are also available.[5]

We follow the same experimental setup as in abstractive summarization. We use the coherence models trained on the WSJ dataset to produce scores that can be used to obtain the pairwise ranking of generated summaries. Based on the ratings provided by Barzilay and Lapata (2008), we can generate the human pairwise rankings.

Results. We present the agreements between the generated human ranking and the systems in Table 6. We observe the same problem as in abstractive summarization that high accuracy in synthetic tasks does not lead to high human agreement in evaluating downstream summarization systems.

3.3 Task-specific Training for Dialog

The global and local discrimination tasks are synthetic, while the MT and summarization coherence evaluation performance may be affected by the difference between the testing and training setup. To control for this, we re-train and test the coherence models on a task-specific setup for next utterance ranking. This task has the advantage of being non-synthetic while providing task specific training data, but also being similar to the synthetic task of insertion, helping us evaluate the generalizability of the coherence model performance.

3.3.1 Next Utterance Ranking

The quality of a dialog depends on various conversational aspects such as engagement, coherence, coverage, conversational depth, and topical diversity (See et al., 2019). Liu et al. (2016) show that commonly used metrics such as BLEU and ROUGE show very weak or no correlation with human judgements. They also suggest using metrics

---

[5]See Appendix for details. Rankings are available at http://homepages.inf.ed.ac.uk/mlap/coherence/

	Train	Dev	Test
Advising dataset			
# of conv.	50,535	500	269
# of coh.-incoh. pairs/conv.	20	99	99
# of total example pairs	10,10,700	49,500	26,631
Ubuntu dataset			
# of conv.	49,387	500	1078
# of coh.-incoh. pairs/conv.	20	99	99
# of total example pairs	9,87,740	49,500	1,06,722

Table 7: Statistics of the refined Advising and Ubuntu datasets for the **utterance ranking** task.

that take dialog context into account. This is particularly important as Sankar et al. (2019) empirically show that current neural dialog systems rarely use conversational history. We therefore propose to evaluate the usefulness of coherence models in dialog systems.

We evaluate the models on the Noetic End-to-End Response Selection Challenge II (NOESIS II), a track in the Dialog System Technology Challenges 8 (DSTC 8) (Kim et al., 2019). In this problem, each example consists of a conversational context $U = (u_1, \dots, u_{|U|})$ and a set of potential utterances (candidates) $C = \{c_1, \dots, c_{|C|}\}$ that may occur next in the dialog; the task is to select the correct next-utterance $r \in C$.

This task is a nice fit for evaluating coherence models, as a good model should rank a coherent dialog higher than an incoherent one. The correct utterance along with the conversational context forms the coherent example $P = (u_1, \dots, u_{|U|}, r)$, while other candidate utterances $c_j \in C$ with the conversational context form the incoherent examples $N = (u_1, \dots, u_{|U|}, c_j)$. This is a considerably harder task as the difference between coherent and incoherent dialog is only the last utterance. We train the coherence models with these coherent (P) and incoherent (N) examples. The trained models give a score for each example based on its coherence. We then use our aforementioned assumption (coherence models should score P higher than N) for the evaluation. This task resembles the (synthetic) *insertion* task (Elsner and Charniak, 2011b) in that the goal here is to find the next correct utterance for the last position.

Setup. We evaluated the coherence models on both datasets of the DSTC8 response selection track, *i.e.,* the Advising and Ubuntu datasets.[6] The

	R@1	R@5	R@10	MRR	Acc.
Advising dataset					
Official Evaluation					
Best	0.564	0.81	0.88	0.68	X
Median	0.14	0.37	0.51	0.26	X
Worst	0.01	0.05	0.09	0.05	X
Coherence Model					
EGRID	0.004	0.03	0.07	0.04	47.16
NEURALEGRID	0.057	0.17	0.23	0.13	56.15
LEXNEUEGRID	0.046	0.17	0.26	0.13	57.66
TRANSMODEL	0.067	0.20	0.30	0.14	**66.62**
UNIFIEDMODEL	0.022	0.06	0.19	0.11	54.33
Ubuntu dataset					
Official Evaluation					
Best	0.761	0.96	0.98	0.85	X
Median	0.55	0.86	0.93	0.68	X
Worst	0.24	0.38	0.46	0.32	X
Coherence Model					
EGRID	0.007	0.05	0.09	0.05	47.48
NEURALEGRID	0.18	0.39	0.49	0.29	73.18
LEXNEUEGRID	0.15	0.31	0.39	0.24	74.39
TRANSMODEL	0.045	0.14	0.26	0.12	70.94
UNIFIEDMODEL	0.035	0.17	0.33	0.13	**74.49**

Table 8: **Utterance ranking** results for different coherence models on Advising and Ubuntu datasets. **R@k** indicates Recall@k, X indicates result not shared.

former contains two-party dialogs that simulate a discussion between a student and an academic advisor, while the latter consists of multi-party conversations extracted from the Ubuntu IRC channel (Kummerfeld et al., 2019).

For a given conversational context, the goal is to select the next utterance from a candidate pool of 100 utterances, which may or may not contain the correct next utterance. We filter the datasets to suit the settings for coherence models. In our refined datasets, we exclude the conversations that have less than 7 or more than 50 utterances in the context. To ensure that we have pairwise coherent and incoherent examples, we only include the conversations that contain the correct next utterance in the candidate pool. Table 7 shows the statistics of our refined datasets for the utterance ranking task.

Results. Table 8 summarizes the results on the refined datasets for the utterance ranking task. In the last column, we report the accuracy for the number of samples in which the coherence models score the positive sample higher than the negative one. All model performances are better than a random baseline, with UNIFIEDMODEL reaching 74.49% on the Ubuntu dataset. Note, however, that because there are 100 negative samples for every positive sample, the accuracies are skewed and not representative of actual task difficulty.

The DSTC8 challenge ranking considers the average of Recall@1, Recall@5, Recall@10 and Mean Reciprocal Rank (MRR). We report both the official evaluation results and the coherence models' performance even though the latter is tested on the refined datasets. From the results, we see that the overall performance of all the coherence models is quite poor. Despite being re-trained on task specific data, we find that coherence model performance in this task is sub-par.

4 Task-specific Training for MT

As a special use case, we report the results of re-training the coherence models using machine translation data for coherence evaluation. The aim is to investigate whether changing the usual training setup, that uses negative documents which are only small variations of the positive documents, might help coherence models learn more useful task-specific features.

Setup. Under the assumption that the reference translations are more coherent at the document level than the system translations, we train the coherence models with the reference text as the positive and the system translation as the negative document, forming a positive-negative document pair. We use the data from WMT-2011 to WMT-2015 for training (28,985 document-pairs), WMT-2016 for development (7,647 document-pairs) and the same test data (WMT-2017 to WMT-2018; 20,680 document-pairs) and study data as used for the previous experiment (§3.2.1).

Results. Table 9 reports the accuracy of the re-trained models and the results of the model ranking comparison against human rankings. Many of the models show improved performance, with the agreements increasing correspondingly. The UNIFIEDMODEL has the highest accuracy improvement by far of 34%, improving from 43.36% to 77.35%. It also has the highest agreement with human rankings at 0.82. We surmise that the model's adaptive pairwise ranking loss along with its additional language model loss boosts its performance on in-domain test data.

5 Discussion

Compared to the downstream tasks of coherence evaluation in MT and extractive and abstractive summarization, the traditional global discrimination task can be considered to be a simpler task

Model	Acc. (%)	AC1 Agr.
EGRID	48.74	0.797
NEURALEGRID	52.58	0.760
LEXNEUEGRID	56.84	0.795
TRANSMODEL	57.65	0.751
UNIFIEDMODEL	**77.35**	**0.828**

Table 9: **Re-trained MT** setting results on WMT2017-2018 data. **Acc**uracies: % of times reference scored higher and **AC1 agr**eements for system translation rankings between annotators and models.

(Elsner and Charniak, 2011b), since the difference between the positive and the negative document is a permutation/re-ordering of the sentences. This may be rendering the models unable to learn features that are useful for downstream applications, which are likely to have other, different kinds of errors.

On the next utterance ranking task, the models fail to generalize and perform quite poorly despite task-specific re-training. The best model performance for the synthetic task of insertion, which is similar, also barely reaches 26% (Elsner and Charniak, 2011b; Nguyen and Joty, 2017). This indicates that the training procedures may not be providing the right setting to learn features that are generic enough to apply to tasks in a harder setup.

In the synthetic tasks, the models' self-supervision comes from distinguishing an original coherent document from its incoherent renderings generated by random permutations of its sentences. This permutation-based self-supervision tries to capture document-level language properties. However, it is quite likely that this is simply a poor approximation of real-world coherence problems. Consider for example that MT systems mostly translate at the sentence-level. Consecutive sentences may lack coherence, but if two system translations of a text are compared, the translations themselves will be in the same order for both. The coherence models are not trained for such (real-world) settings.

Another possibility is that outputs from downstream tasks have different error distributions that are captured to varying degrees by different models, since they are originally designed based on synthetic tasks. That is, models that perform very well on the permutation task might be overfitting on this task, and therefore failing to find coherence issues that are more subtle than shuffled text. Thus, we conclude that the current self-supervision for coherence modeling is not suitable for downstream

coherence problems.

When re-trained on machine translation data, most of the model performances improve, implying that a different training setting may be required to make the models applicable to actual downstream tasks. This is not apparent from the evaluation results that are usually reported, which show performances crossing the 90% mark.

Elsner and Charniak (2011a) show a similar lack of generalizability and applicability of coherence models to the downstream task of chat disentanglement. Our results suggest that despite nearly a decade of research since, the standard training and testing paradigm for coherence modeling continues to be inadequate in its capability to generalize to real-world use-cases and even to similar task settings, and also fails in being indicative of real-world task performance.

6   Conclusions

We benchmark the performance of representative traditional and neural coherence models on standard synthetic discrimination tasks, and contrast this with their performance on various downstream application tasks in NLP. We show that higher accuracies on synthetic tasks do not translate into better performance on downstream tasks. We demonstrate this for real-world tasks like MT and summarization coherence evaluation, and next utterance ranking. Our results signal a need for change in the way coherence models are typically trained and evaluated.

Other downstream applications like coherence evaluation of language model generated text and tasks such as chat disentanglement are also good candidates for testing coherence models. It would be worthwhile to build a coherence testset that is independent of the training tasks and similar to downstream applications, which could be used by the community to test the generalization ability of their models. In future work, we also hope to investigate the possible training scenarios that will result in more generalizable coherence models which can be used for evaluating downstream tasks.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 141–

148, Ann Arbor, Michigan. Association for Computational Linguistics.

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

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL 2004: Main Proceedings*, pages 113–120, Boston, Massachusetts, USA. Association for Computational Linguistics.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *Prague Bull. Math. Linguistics*, 108:109–120.

Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. 2018. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675, New Orleans, Louisiana. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 436–443, Rochester, New York. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2011a. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1179–1189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2011b. 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*, HLT '11, pages 125–129, Portland, Oregon. Association for Computational Linguistics.

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

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

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Michael Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*, chapter xx. Longman, London.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan R. Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, T. M. Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *ArXiv*, abs/1803.05567.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue.

Seokhwan Kim, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, Minlie Huang, Luis Lastras, Jonathan K. Kummerfeld, Walter S. Lasecki, Chiori Hori, Anoop Cherian, Tim K. Marks, Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, and Raghav Gupta. 2019. The eighth dialog system technology challenge.

Wojciech Kryściński, Nitish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros Polymenakos, and Walter S. Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *EMNLP*.

Jiwei Li and Eduard Hovy. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2039–2048, Doha, Qatar. Association for Computational Linguistics.

Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.

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*, HLT '11, pages 997–1006, Portland, Oregon. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. 2017. Generative adversarial network for abstractive text summarization. *ArXiv*, abs/1711.09357.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. 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*, EMNLP-CoNLL '12, pages 1157–1168, Stroudsburg, PA, USA. Association for Computational Linguistics.

W. Mann and S. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281.

Mohsen Mesgar and Michael Strube. 2018. A neural local coherence model for text quality assessment. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4328–4339, Brussels, Belgium. Association for Computational Linguistics.

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

Muhammad Tasnim Mohiuddin, Shafiq Joty, and Dat Tien Nguyen. 2018. Coherence modeling of asynchronous conversations: A neural entity grid approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 558–568, Melbourne, Australia. Association for Computational Linguistics.

Han-Cheol Moon, Tasnim Mohiuddin, Shafiq Joty, and Chi Xu. 2019. A unified neural coherence model. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP'19, pages xx—-xx, Hong Kong. ACL.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GuÌ‡lçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Ani Nenkova and Kathleen McKeown. 2011. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.

Dat Nguyen and Shafiq Joty. 2017. A neural local coherence model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1320–1330. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *ArXiv*, abs/1705.04304.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT'18, pages 2227–2237, New Orleans, LA, USA.

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. Association for Computational Linguistics.

M. Popel, M. Tomková, J. Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondrej Bojar, and Z. Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019. An entity-driven framework for abstractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3271–3282.

Karin Sim Smith, Wilker Aziz, and Lucia Specia. 2015. A proposal for a coherence corpus in machine translation. In *DiscoMT@EMNLP*.

Karin Sim Smith, Wilker Aziz, and Lucia Specia. 2016. The trouble with machine translation coherence. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 178–189.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 803–810, Sydney, Australia. Association for Computational Linguistics.

B. Webber. 2004. D-LTAG: Extending Lexicalized TAG to Discourse. *Cognitive Science*, 28(5):751–779.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Florence, Italy. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

A   Appendix

A.1   Human Study Interface for Abstractive Summarization

We show the interface of human study for abstractive summarization in Figure 2.

A.2   Human Study for Extractive Summarization

We briefly describe the human study for extractive summarization. The human study was conducted by Barzilay and Lapata (2008). Coherence ratings for summaries were collected during an elicitation study by 177 unpaid native speakers of English. The annotators were asked to use a seven point-scale to rate each summary based on how coherent the summaries were without having seen the source texts. The ratings (approximately 23 per summary) given by the subjects were averaged to provide a final rating score between 1 and 7 for each summary.

Please sort the summaries according to the coherence aspect.
The summaries are given as follows:

(1): eric garner 's family and other members of families united for justice will attend gray 's funeral . gray was arrested april 12 and died a week later from a severe spinal cord injury . three white house officials will also attend gray 's funeral .

(2): Freddie Gray , 25 , died in police custody 15 days ago after he was arrested on a weapons charge . His family said his voice box was crushed and his neck snapped before he slipped into a coma . Hundreds of protesters peacefully rallied on the streets of Baltimore on Saturday against the alleged police role in Gray 's death . " There are those who will tell you

Coherence Ranking:

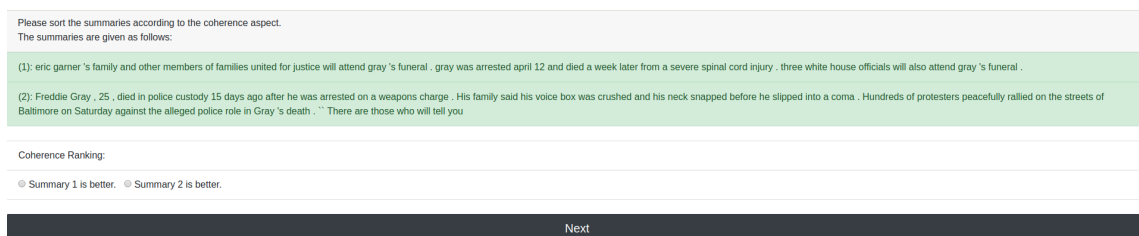○ Summary 1 is better.   ○ Summary 2 is better.

Next

Figure 2: Human Study Interface for Abstractive Summarization