Analyzing Neural Discourse Coherence Models

Youmna Farag[1,2]   **Josef Valvoda**[1]   **Helen Yannakoudakis**[3]   **Ted Briscoe**[1,2]

[1]Department of Computer Science and Technology, University of Cambridge, United Kingdom
[2]The ALTA Institute, Cambridge, United Kingdom
{youmna.farag,jv406,ted.briscoe}@cl.cam.ac.uk
[3]Department of Informatics, King's College London, United Kingdom
helen.yannakoudakis@kcl.ac.uk

Abstract

In this work, we systematically investigate how well current models of coherence can capture aspects of text implicated in discourse organisation. We devise two datasets of various linguistic alterations that undermine coherence and test model sensitivity to changes in syntax and semantics. We furthermore probe discourse embedding space and examine the knowledge that is encoded in representations of coherence. We hope this study shall provide further insight into how to frame the task and improve models of coherence assessment further. Finally, we make our datasets publicly available as a resource for researchers to use to test discourse coherence models.

1 Introduction

Coherence refers to the properties of a text that indicate how meaningful (sub-)sentential constituents are connected to convey document-level meaning. Different theories have been proposed to describe the properties that contribute to discourse coherence and some have been integrated with computational models for empirical evaluation. A popular approach is the entity-based model which hypothesizes that coherence can be assessed in terms of the distribution of and transitions between entities in a text – by constructing an entity-grid (Egrid) representation (Barzilay and Lapata, 2005, 2008), building on Centering Theory (Grosz et al., 1995). Subsequent work has adapted and further extended Egrid representations (Filippova and Strube, 2007; Burstein et al., 2010; Elsner and Charniak, 2011; Guinaudeau and Strube, 2013). Other research has focused on syntactic patterns that co-occur in text (Louis and Nenkova, 2012) or semantic relatedness between sentences (Lapata and Barzilay, 2005; Soricut and Marcu, 2006; Somasundaran et al., 2014) as key aspects of coherence modeling. There have also been attempts to model coherence

by identifying rhetorical relations that connect textual units (Mann and Thompson, 1988; Lin et al., 2011; Feng et al., 2014) or capturing topic shifts via Hidden Markov Models (HMM, Barzilay and Lee, 2004). Other work has combined approaches to study whether they are complementary (Elsner et al., 2007; Feng et al., 2014). More recently, neural networks have been used to model coherence. Some models utilize structured representations of text (e.g. Egrid representations, Tien Nguyen and Joty, 2017; Joty et al., 2018) and others operate on unstructured text, taking advantage of neural models' ability to learn useful representations for the task (Li and Jurafsky, 2017; Logeswaran et al., 2018; Farag and Yannakoudakis, 2019; Xu et al., 2019; Moon et al., 2019).

Coherence has typically been assessed by a model's ability to rank a well-organized document higher than its noisy counterparts created by corrupting sentence order in the original document (*binary discrimination task*), and neural models have achieved remarkable accuracy on this task. Recent efforts have targeted additional tasks such as recovering the correct sentence order (Logeswaran et al., 2018; Cui et al., 2018), evaluating on realistic data (Lai and Tetreault, 2018; Farag and Yannakoudakis, 2019) and focusing on open-domain models of coherence (Li and Jurafsky, 2017; Xu et al., 2019). However, less attention has been directed to investigating and analyzing the properties of coherence that current models can capture, nor what knowledge is encoded in their representations and how it might relate to aspects of coherence.

In this work, we systematically examine what properties of discourse coherence current coherence models can capture. We devise two datasets that exhibit various kinds of incoherence and analyze model ability to capture syntactic and semantic aspects of text implicated in discourse organisation. We furthermore investigate a set of probing tasks to

better understand the information that is encoded in their representations and how it might relate to aspects of coherence. We hope this study shall provide further insight into how to frame the task and improve models of coherence assessment further. Finally, we release our evaluation datasets as a resource for the community to use to test discourse coherence models.[1]

2 Neural Coherence Models

We experiment with a number of existing and state-of-the-art neural approaches to coherence assessment, that have publicly available implementations, and present details of the models below. Across all the BERT-based models, we use bert-large-uncased and layer 16 following Liu et al. (2019) and Hewitt and Manning (2019).

Multi-task learning (**MTL**, Farag and Yannakoudakis, 2019): The model applies a Bi-LSTM on input GloVe word embeddings (Pennington et al., 2014) followed by attention to build sentence representations; then builds a second Bi-LSTM with attention to compose a document vector. A linear operation followed by a sigmoid function is applied to the document representation to predict an overall coherence score as the main objective. Inspired by the Egrid approaches, the model is also optimized to predict the grammatical roles of the input words at the bottom layer of the network as an auxiliary task.

MTL with BERT embeddings (**MTL**$_{bert}$): We replicate the previous MTL model but now use BERT embeddings (Devlin et al., 2019) to initialize the input words.

Single-task learning (**STL**, Farag and Yannakoudakis, 2019): This model has the same architecture as MTL but only performs the coherence prediction task, excluding the grammatical role auxiliary objective.

STL with BERT (**STL**$_{bert}$): This is the same as STL but uses BERT embeddings.

Local Coherence Discriminator with Language modeling (**LCD**$_{rnnlm}$, Xu et al., 2019): The model generates sentence representations via an RNN language model, where word embeddings are initialized using GloVe. It then generates a representation for two consecutive sentences via concatenating the output of a set of linear transformations applied to the two sentences: concatenation, element-wise dif-

ference, element-wise product and absolute value of element-wise difference. This representation is fed to an MLP layer to predict a *local* coherence score.[2] The overall coherence of a document is the average of its local scores.

LCD with BERT (**LCD**$_{bert}$): We create a variant of the LCD$_{rnnlm}$ model where instead of using an RNN language model encoder, we encode each sentence as the average BERT vectors of the words it contains. Everything else remains the same.

Local Coherence (LC, Li and Jurafsky, 2017): The model generates sentence vectors via an LSTM over GloVe-initialized word embeddings; then a window approach is applied over adjacent sentences to get embeddings of groups of sentences and predict local coherence scores. The final document score is calculated by averaging its local scores.

Egrid CNN (Egrid$_{cnn}$, Tien Nguyen and Joty, 2017): The model applies a CNN over Egrid representations across groups of consecutive sentences; the CNN slides multiple filters of weights to extract feature maps that represent high-level entity-transition features, followed by a max pooling function to focus on the important features. Furthermore, additional entity-related features are integrated such as salience, proper mentions and named entity type.

3 Binary Discrimination Task

Binary discrimination is a typical approach to assessing neural coherence models where a well-organized document should be ranked higher than its permuted counterparts created by corrupting sentence order. Following previous work, we train and test[3] the coherence models on the WSJ[4] and evaluate them using Pairwise Ranking Accuracy (PRA), which is calculated based on the fraction of correct pairwise rankings between a coherent document and its incoherent counterparts.

In Table 1, we present the performance of all coherence models. The high accuracy of the models demonstrates their efficacy for the task of selecting a maximally coherent sentence order from a set of candidate permutations. We note that the LCD and

[2] Gold local scores $\in \{0, 1\}$ represent whether a sequence of two sentences is coherent (i.e. extracted from a coherent document) or not (i.e. created via negative sampling).

[3] All models are run 5 times and the test predictions are averaged across the runs.

[4] We use the same train and test splits as Tien Nguyen and Joty (2017) and the same test set permuted counterparts as Farag and Yannakoudakis (2019).

MTL	MTL$_{bert}$	STL	STL$_{bert}$	LCD$_{rnnlm}$	LCD$_{bert}$	LC	Egrid$_{cnn}$
93.2	96.1	87.7	95.4	94.5	**97.1**	74.1	87.6

Table 1: PRA results of coherence models based on the binary discrimination task on the WSJ.

MTL BERT variants achieve a new state-of-the-art on the WSJ. The remarkable accuracy on this task may render this problem fully solved.

Herein, we seek to investigate how well these models of coherence can capture aspects of text implicated in discourse organisation. We devise a set of datasets and systematically test model susceptibility to syntactic or semantic changes.

4 Cloze Coherence (CC) Dataset

We compile a large-scale dataset, to which we refer as Cloze Coherence (CC), of coherent and incoherent examples, where the former are intact well-written texts while the latter are the result of applying syntactic or semantic perturbations to the coherent ones.

4.1 Coherent examples

For the sake of specifically testing for coherence, we avoid complex linguistic structures. Specifically, we focus on coherent examples that consist of two short sentences that are coreferential and exhibit a rhetorical relation (such properties can be manipulated to create incoherent counterparts). Furthermore, we focus on examples that are self-contained, meaning that they do not reference or rely on an outer context to be interpreted. We find that narrative texts are good candidates to satisfy these criteria and therefore create our coherent examples from the ROCStories Cloze dataset[5] (Mostafazadeh et al., 2016).

ROCStories Cloze contains short stories of 5 sentences manifesting a sequence of causal or temporal events that have a shared protagonist. A story usually starts by introducing a protagonist in the first sentence, then subsequent sentences describe events that happen to them in a logical / rhetorically plausible manner. The dataset was designed for commonsense reasoning by testing the ability of machine learning models to select a plausible ending for the story out of two alternative endings. Here, our main aim is to challenge the models and investigate whether they truly understand intersentential relations and coherence-related features. We specifically utilize the first two sentences in the

stories to compose the coherent examples in our dataset.[6]

Selecting the first two sentences helps make the examples self-contained since there is no preceding context to refer to, and no cataphoric relations to consequent sentences. Regarding rhetorical relations in these sentences, Mostafazadeh et al. (2016) conducted a temporal analysis to investigate the logical order of the events presented in a story, demonstrating, among others, that the first and second sentences in the stories are presented in a commonsensical temporal manner with logical links between them. In order to examine coreferential relations between the two sentences in each extracted pair, we gather a set of statistics. We adopt a heuristic approach[7] by simply counting the number of second sentences that contain at least one third person pronoun (either personal or possessive) and find that they constitute 80% of the examples.[8] Third person pronouns anaphorically refers to preceding items in text, which could occur in the same sentence or the previous one (i.e., the first sentence). We, therefore, randomly select, and manually inspect, 500 examples that contain third person pronouns in their second sentence and find that in 95% of them the referenced entity appears in the first sentence. Furthermore, third person pronouns are not the only coreferential relations in the examples. For instance, we find that 90% of the second sentences contain a personal or possessive pronoun (whether it is first, second or third person), which could also signal coreference, e.g., '*I was walking to school. Since I wasn't looking at my feet I stepped on a rock.*' There are also other coreferential devices such as: demonstrative references (e.g., 'this' and 'there'), 'the' + noun, proper names or nominal substitutions (e.g., 'one' or 'ones') to name a few (Halliday and Hasan., 1976), so the true proportion of coreferential pairs will be higher. Table 2 presents examples of different referential relations in our dataset.

---

[5] https://www.cs.rochester.edu/nlp/rocstories/

[6] We use NLTK for word tokenization; sentence boundaries are already marked in the stories.

[7] We initially used the spaCy and Stanford coreference resolution systems (Clark and Manning, 2016), but found their performance unreliable for the purposes of this experiment after manual inspection.

[8] If we exclude 'it' the percentage becomes 76%.

Type of reference	Example
Pronominal Reference	Rich was a musician. <u>He</u> made a few hit songs.
Proper Name	Dan's parents were overweight. <u>Dan</u> was overweight as well.
Nominal Substitution	My dog hates his treats. I decided to go buy some new <u>ones</u>.
Demonstrative Reference	My daughter wants to take her toddler to the Enchanted Village. <u>This</u> is a puppet show featuring early 20th century figurines.

Table 2: Examples of first two sentences extracted from the ROCStories Cloze dataset with different referential types (referring word underlined).

We use the same train/dev/test splits provided with ROCStories Cloze but only keep the first two sentences in each story. We exclude cases with erroneous sentence boundaries,[9] yielding $97,903$ examples for training, $1,871$ for development, and $1,871$ for testing, and a training vocabulary size of $29,596$ tokens. Each instance in our dataset contains two sentences that represent a coherent pair.

4.2 Incoherent examples

To assess model susceptibility to syntactic or semantic alterations, we construct incoherent examples by applying two different transformations to each coherent pair resulting in two different sets of data.

cloze_swap We create incoherent examples by swapping the two sentences in a coherent pair. This mostly breaks the coreference relation between them and/or the rhetorical relation (e.g. temporal or causal) by reversing the event sequence. The dataset, referred to as cloze_swap, is balanced, i.e., the number of incoherent examples is the same as the number of the coherent ones above. The way cloze_swap is created corrupts the syntactic patterns that co-occur in coherent texts (e.g. S $\rightarrow$ NP-SBJ VP | NP-SBJ $\rightarrow$ PRP) as demonstrated by Louis and Nenkova (2012).

cloze_rand Here we create incoherent examples by keeping the first sentence of a coherent pair intact and replacing the second with a randomly selected second sentence from (the same split of) our set of coherent examples. This dataset, referred to as cloze_rand, is also balanced (for each coherent pair, we compose one incoherent counterpart), and constitutes examples with changed semantics but with the main syntactic pattern intact. As the randomly-created pair may still be coherent, we address this by: 1) constraining random selection of the second sentence to not begin with the same

word as the second sentence in the original pair, or with the pronoun 'he' if the original starts with 'she', and vice-versa[10] (we note 70% of the second sentences in ROCStories Cloze start with a pronoun); 2) using human evaluation to further assess the validity of this data and get an estimate of upper-bound performance on the task. Specifically, we randomly select 100 coherent sentence pairs from our test split along with their own incoherent counterparts and ask two annotators (who are not authors of this paper), with high English proficiency levels, to *rank* each set of coherent–incoherent examples based on which one they considered to be more coherent and plausible. The average PRA of the annotators is 94.5%.

Table 3 shows examples from cloze_swap and cloze_rand. As our datasets are balanced (one incoherent counterpart per coherent pair), we have a total number of $195,806$, $3,742$ and $3,742$ instances in the train, dev, and test splits respectively for each cloze dataset (cloze_swap and cloze_rand have the same coherent examples, and the same number of coherent and incoherent examples).

We note that the gold labels in this data are not to be interpreted as (overall) binary indicators of coherence. We rather use these to test model performance using PRA, i.e. we only compare a coherent pair with *its own* incoherent counterpart.

5 Controlled Linguistic Alterations (CLA) Dataset

In order to further understand the properties of coherence that current coherence models capture, we manually construct a dataset of controlled sets of linguistic changes. We first identify a set of coherent, well-written texts of two consecutive sentences from business and financial articles in the BBC, the Independent and Financial Times (this allows us to stay in the same domain as the one used for training the models – the WSJ). We focus on sentence pairs where the subject of the first sentence is pronom-

---

[9] The training stories are in CSV format (separating sentences by comma delimiters) and we parse them using the Python CSV parser. We exclude the stories where the parser fails to detect 5 sentences.

[10] We do not find instances of 'they' as a third-person singular pronoun.

Coherent example	Incoherent example from cloze_swap	Incoherent example from cloze_rand
Tyrese joined a new gym.The membership allows him to work out for a year.	The membership allows him to work out for a year. Tyrese joined a new gym.	Tyrese joined a new gym. As children they hated being dressed alike.
Jasmine doesn't know how to play the guitar. She asked her dad to take her to guitar class.	She asked her dad to take her to guitar class. Jasmine doesn't know how to play the guitar.	Jasmine doesn't know how to play the guitar. May thought her milk was no good.
I wanted to play an old game one day. When I looked in the game's case the CD was missing.	When I looked in the game's case the CD was missing. I wanted to play an old game one day.	I wanted to play an old game one day. Jason pressed the buzzer since he knew the answer.

Table 3: Examples of coherent and incoherent pairs from the cloze_swap and cloze_rand datasets.

Original	A government paper on Monday found UK and EU firms would be faced with a "a significant new and ongoing administrative burden" in the event of a no-deal Brexit. It found large firms importing and exporting at scale would need to fill in forms taking one hour 45 minutes on average and cost £28 per form for each load imported.
Swap	It found large firms importing and exporting at scale would need to fill in forms taking one hour 45 minutes on average and cost £28 per form for each load imported. A government paper on Monday found UK and EU firms would be faced with a "a significant new and ongoing administrative burden" in the event of a no-deal Brexit.
Random	1- A government paper on Monday found UK and EU firms would be faced with a "a significant new and ongoing administrative burden" in the event of a no-deal Brexit. She spent over a decade at Swiss investment bank UBS before joining the UK Treasury's council of economic advisers in 1999. 2- Lady Vadera was born in Uganda and moved to the UK as a teenager. It found large firms importing and exporting at scale would need to fill in forms taking one hour 45 minutes on average and cost £28 per form for each load imported.
Lexical Substitution	The paper found large firms importing and exporting at scale would need to fill in forms taking one hour 45 minutes on average and cost £28 per form for each load imported. A government paper on Monday found UK and EU firms would be faced with a "a significant new and ongoing administrative burden" in the event of a no-deal Brexit.
Prefix Insertion	More Specifically, it found large firms importing and exporting at scale would need to fill in forms taking one hour 45 minutes on average and cost £28 per form for each load imported. A government paper on Monday found UK and EU firms would be faced with a "a significant new and ongoing administrative burden" in the event of a no-deal Brexit.
Lexical Perturbations	A government paper on Monday found UK and EU firms would be faced with a "a significant new and ongoing administrative burden" in the event of a no-deal Brexit. It found large firms importing and exporting at scale would need to fill in cups taking one hour 45 minutes on average and cost £28 per cup for each load imported.
Corrupt Pronoun	A government paper on Monday found UK and EU firms would be faced with a "a significant new and ongoing administrative burden" in the event of a no-deal Brexit. He found large firms importing and exporting at scale would need to fill in forms taking one hour 45 minutes on average and cost £28 per form for each load imported.

Table 4: Examples from our manually constructed CLA dataset. For 'Random' we create two incoherent instances: one where the first sentence is unchanged and the second is randomly selected (1-); and another where the first sentence is randomly selected and the second is kept intact (2-).

inalized in the second, and the second sentence begins with this pronoun. We select the examples so that they are self-contained and do not reference an outer context. We then manually create incoherent counterparts by modifying the coherent examples in a constrained way in order to systematically examine model performance. Specifically, we apply the following sets of perturbations to our set of coherent sentence pairs, examples of which are presented in Table 4.

Swap. We simply swap the two sentences.

Random. We keep the first sentence intact and select a second sentence randomly from our set of coherent examples. We constrain the selection so that the subject pronoun is different from the subject pronoun in the original sentence.[11] We also create another random pair with the same constraint but now changing the first sentence. Thus each original coherent example has two incoherent counterparts.

Lexical Substitution. We swap the two sentences in a coherent pair but replace the subject pronoun in the second sentence with *the + a general noun* that substitutes the subject in the first sentence (e.g.

the company, the woman, etc.).

Prefix Insertion. We analyze the WSJ training data and find that the average number of times the first sentence in a document starts with a pronoun is 0.02 (and never with 'he' or 'she') which is significantly less than the average number of times a sentence starts with a pronoun (regardless of its position) which is 0.07. This difference is not maintained in the randomly ordered documents in the WSJ training set and so this might give a signal to the models to detect that a swapped pair that starts with a pronoun is less coherent. To see if such positional information plays a role in model prediction, we insert a phrase, before the subject pronoun after swapping the sentences, that doesn't change the propositional content (e.g. 'More specifically', 'However', etc.). We can then observe whether this insertion will change the prediction of the model.

Lexical Perturbation. We investigate the robustness of the models to minor lexical changes that result in incoherent meaning, by replacing one word in either of the two sentences (if the word is repeated, we change that too). We choose a replacement word from the training vocabulary of the WSJ with the same part-of-speech tag. For example, in Table 4 'form' is replaced with 'cup' and 'forms' with 'cups'.

---

[11] We also take into account that some subjects could be referred to by 'he', 'she' or 'they' and thus factor that into the selection.

Corrupt Pronoun. We replace the subject pronoun in the second sentence with another pronoun that cannot reference anything in the first sentence. With this method, we test whether the models are capable of resolving coreferences or just rely on syntactic patterns.

Our dataset contains a total of 240 examples of coherent and incoherent pairs of sentences (30 coherent examples and 210 incoherent counterparts). Our constrained set of modifications ensures that all coherent examples are more coherent than any of the incoherent counterparts in the data.

6 Experiments

Table 5 (top) presents the PRA performance of the models trained on the WSJ (Section 3) when they are evaluated on the test sets of the CC datasets (rows 'cloze_swap' and 'cloze_rand'). We find that, overall, models are good at detecting syntactic alterations (cloze_swap; PRA ranging from 69.3 to 84.6) even though the test data is from a domain different than the training one. However, most models perform poorly on semantic alterations (cloze_rand; PRA ranging from 48.5 to 54.5), the only exception being LCD$_{bert}$ that achieves a PRA of 71. Specifically, models that use RNN-based sentence encoders (the first six models), even when initialised with BERT, or apply a CNN to capture entity transitions fall short in capturing semantic changes despite the fact that cloze_rand is from the same domain as cloze_swap. In contrast, LCD$_{bert}$ is more capable of detecting semantic changes where the model builds sentence representations by averaging BERT vectors then applies a set of linear transformations to increase its expressive power, surpassing its RNN-based counterpart (LCD_rnnlm) with 16.5% on cloze_rand. Additionally, across models, we observe that the use of contextualized (BERT) embeddings consistently improves performance on both cloze tasks, although performance on semantic alterations remains close to random.

We investigate domain shift effects and fine-tune the WSJ-trained models on each of the cloze_swap and cloze_rand training sets (Section 4) and re-evaluate performance on the respective test sets. Specifically, we use an MLP layer over the models' pre-prediction representation, followed by sigmoid non-linearity. The models are optimized using the mean squared error between the gold labels (0 or 1) and the predicted scores.[12] In this setup, only the MLP layer is fine-tuned and not the whole coherence model which allows us to create a fast efficient evaluation framework that can be applied as a further examination step after coherence models are developed and tuned on their respective datasets, instead of training the models from scratch. The results of the fine-tuned models are presented in Table 5 (CC; rows 'fine-tuned'). Although we can see that there is some domain effect, we nevertheless find that the results confirm our earlier observation: performance on semantic alterations is, overall, poor, in contrast to syntactic ones (cloze_swap).

In Table 5 (bottom), we can observe model performance (PRA) on our constrained set of manually devised examples (CLA). Again, we observe a similar result: across RNN-based models, performance is particularly low on random examples, which suggests that they struggle to detect topical or rhetorical shifts and unresolved references if the main syntactic pattern is maintained. The exception is LCD_bert which is again the best performing model (PRA 78.3).

We furthermore observe that now Egrid$_{cnn}$ is the second best model on CLA Random (PRA 71.6). A sparser entity grid where entities in the two sentences are different allows the model to detect such cases (e.g. in the example in Table 4, 'firms' is mentioned in the two sentences, while in the two random examples, it is only mentioned in one). However, its substantial difference in PRA on CLA Random compared to cloze_rand (53.4) suggests that the lower performance observed in the latter is due to domain shift effects, something which we do not observe (to the same extent) with LCD$_{bert}$. Regarding the CLA Swap results, we can again confirm models' capability of detecting corrupted syntactic constructions. We furthermore observe that they are able to maintain good performance in the cases where a prefix is inserted ('Prefix Insertion') or the subject pronoun is substituted with a lexical item ('Lexical Substitution'). This suggests that they can capture the relevant syntactic patterns and do not rely solely on positional features.

Performance is overall low on lexical perturbations and corrupt pronouns which suggests that the models are not sensitive to minor lexical changes even if they result in implausible meaning and they also struggle to resolve pronominal references.

---

[12]We use Adam (Kingma and Ba, 2015), batch size 64, L2 regularization, and a learnable penalty rate (search space {0.00001, 0.0001, 0.001, 0.01}). We use early stopping and stop training if PRA does not improve on the dev set over 5 epochs (max epochs 200). MLP hidden unit size is 100.

	Dataset	# comparisons	Models							
			MTL	MTL_{bert}	STL	STL_{bert}	LC	LCD_{rnnlm}	LCD_{bert}	$Egrid_{cnn}$
CC	cloze_swap	1,871	69.3	73.5	74.2	75.3	70.7	74.5	75.4	**84.6**
	fine-tuned	1,871	88.8	88.5	83.5	84.7	76.3	88.4	**96.7**	88.1
	cloze_rand	1,871	51.3	53.3	48.5	52.5	50.5	54.5	**71.0**	53.4
	fine-tuned	1,871	65.7	54.2	53.7	56.1	51.3	65.2	**94.8**	68.8
CLA	Swap	30	90.0	**93.3**	83.3	90.0	80.0	**93.3**	86.6	83.3
	Random	60	56.6	45.0	50.0	51.6	51.6	61.6	**78.3**	71.6
	Lexical Substitution	30	83.3	**93.3**	80.0	90.0	86.6	83.3	86.6	76.6
	Prefix Insertion	30	83.3	**96.6**	76.6	90.0	76.6	86.6	93.3	80.0
	Lexical Perturbations	30	56.6	46.6	46.6	63.3	50.0	53.3	**80.0**	53.3
	Corrupt Pronoun	30	70.0	53.3	63.3	63.3	53.3	60.0	**76.6**	56.6
	All data	210	70.9	67.6	64.2	71.4	64.2	71.4	**82.8**	70.4
	All data (TPRA)	6,300	69.9	71.3	61.8	71.6	66.0	69.1	**72.2**	65.8

Table 5: PRA performance on the CLA (bottom) and CC datasets (top; 'fine-tuned' shows results for models tuned on the respective cloze training sets).

Task	Models								Human
	MTL	MTL_{bert}	STL	STL_{bert}	LC	LCD_{rnnlm}	LCD_{bert}	Best from Conneau et al. (2018)	
SubjNum	64.9	75.4	62.2	71.5	52.7	71.2	**88.0**	95.1 (Seq2Tree)	88.0
ObjNum	64.5	72.1	61.1	70.7	54.5	65.0	**86.5**	95.1 (Seq2Tree)	86.5
CoordInv	58.5	63.4	53.0	63.7	53.0	56.6	**78.4**	76.2 (NMT En-De)	85.0
CorruptAgr	53.2	69.7	57.7	68.6	52.2	64.2	**94.3**	-	-

Table 6: Classification accuracy on probing tasks. 'Human' shows the human upper bound on the task.

However, the exception is LCD_{bert} (with PRA 80 on lexical perturbations and 76.6 on corrupt pronoun) suggesting a better ability at capturing semantics and resolving references.

Across all six CLA datasets ('All data'; Table 5), we find that, overall, LCD_{bert} is the top performing model (average PRA). The 'All data' row reports the result of comparing a coherent example against its incoherent counterparts across the different alterations (i.e., in Table 4, the original example is compared against all the examples in the table and this is applied to all the original examples in the dataset). If we furthermore compare all the coherent examples against the incoherent ones in the whole dataset (rather than against their own incoherent counterparts), we find that a similar performance pattern is maintained (row 'All data (TPRA)', i.e., all data Total Pairwise Ranking Accuracy).

7 Probing Coherence Embedding Space

Inspired by previous work (Conneau et al., 2018), and to better understand the information that is encoded in the representations of coherence models, we investigate probing tasks that can capture coherence-related features.

We experiment with the following set of sentence-level tasks that are relevant to discourse coherence: 1) the subject number (SubjNum) task that detects the number of the subject of the main clause; 2) the object number (ObjNum) task that detects the number of the direct object of the main clause; 3) the coordination inversion (CoordInv) task that contains sentences consisting of two coordinate clauses, where the two clauses are inverted in half of the sentences and kept intact in the other half (the task is to detect whether a sentence is modified or not); 4) the corrupt agreement (CorruptAgr) task where sentences are corrupted by inverting the verb number (the task is to identify corrupted sentences).

Tasks 1, 2 and 4 align with Centering theory as they probe for subject and object relevant information; the theory suggests that subject and object roles are indicators of entity salience. On the other hand, task 3 tests whether the models can capture intra-sentential coherence. For these tasks, we use the datasets from Conneau et al. (2018) (tasks 1,2 and 3) and Linzen et al. (2016) (task 4).

Probing model We adopt the SentEval framework of Conneau et al. (2018). Our probing model consists of an MLP layer over model sentence representations, followed by sigmoid non-linearity. We use the same training parameters as Conneau et al. (2018).[13]

Results Table 6 presents the results.[14] Overall, we observe that models are better at detecting Sub-

---

[13] https://github.com/facebookresearch/SentEval

[14] $Egrid_{cnn}$ is based on entity transitions across sentences and therefore we cannot probe sentence representations.

jNum, and ObjNum (accuracy of at least 61% for all models except LC which is the odd one out) compared to CorruptAgr and CoordInv, with the last two being particularly challenging for most models (minimum accuracy of 53% excluding LC). For SubjNum and ObjNum the models can find hints in words other than the target word (as the majority of nouns in a sentence tend to have the same number, with 75.9% of SubjNum test sentences and 78.7% of ObjNum ones containing nouns of the same number in the same sentence (Conneau et al., 2018)). On the other hand, CorruptAgr examples are longer and with more syntactic variations and require the models to detect the dependency between verbs and their subjects. CoordInv is also a difficult task for the models particularly since they are pre-trained on the WSJ to focus on the order of sentences, not clauses.

Across all tasks, we find that LCD_{bert} achieves the best performance, outperforming all other approaches. We note, however, that LCD_{bert} does not fine-tune its sentence representations during coherence training in the WSJ but they are rather fixed and based on the average of BERT-based word embeddings (Section 2). This means the probing model fine-tunes averaged BERT-based word embeddings rather than actual sentence parameters from the LCD coherence model. Therefore, the level of performance observed is not representative of the maximum performance coherence models can achieve on these tasks.[15] We surmise that the comparatively lower performance observed with MTL_{bert} and STL_{bert} (whose sentence representations are fine-tuned during coherence training) is due to their coherence training objective. The models are optimized on the binary discrimination task, i.e. learning to rank a well-organized document higher than its permuted counterparts. This is an overly simplistic approach to coherence modeling that may be making models (and their representations) more susceptible to losing useful linguistic information. Having said that, though, MTL_{bert}, that has direct training signal with respect to the words' grammatical roles, is able to alleviate this issue to an extent and is the next best performing model on SubjNum, ObjNum and CorruptAgr.

Across tasks, LC is the odd one out, and the worst performing model. This can be explained partly by its comparatively lower performance on

the simpler binary discrimination task (Table 1) and partly by the simplicity of the approach: LC utilizes no attention mechanism as the MTL and STL family of models do, nor has expressive enough transformations as LCD_{rnnlm} does.

8 Discussion

Our evaluation experiments on two coherence datasets reveal that RNN- or EGrid-based coherence models are able to detect syntactic alterations that undermine coherence, but are less effecient at detecting semantic ones even after fine-tuning on the latter. We furthermore find that they particularly struggle with recognizing minor lexical changes even if they result in implausible meaning and resolving pronominal references. On the other hand, these models are particularly good at detecting cases where a prefix is inserted or the subject pronoun is substituted with a lexical item, suggesting that they are capable of capturing the relevant syntactic patterns and do not solely rely on positional features. We find that the best performing model overall is LCD_{bert} which does not use an RNN sentence encoder but rather builds sentence representations by averaging BERT embeddings then utilizes a number of linear transformations over adjacent sentences to facilitate learning richer representations.

Our probing experiments reveal that models are better at encoding information regarding subject and object number followed by verb number (CorruptAgr). These probing tasks align with Centering theory as they probe for subject and object relevant information. The task that tests for knowledge on coordination inversion is the lowest performing one overall, suggesting that there is little capacity at capturing information related to intra-sentential coherence. Excluding LCD_{bert}, MTL_{bert} is the best performing model; nevertheless, there is still scope for substantial improvement across all probing tasks and particularly on CoordInv and CorruptAgr.

9 Conclusion

We systematically studied how well current models of coherence can capture aspects of text implicated in discourse organisation. We devised datasets of various kinds of incoherence and examined model susceptibility to syntactic and semantic alterations. Our results demonstrate the models are robust with respect to corrupted syntactic patterns,

---

[15]Nevertheless, we observe that LCD_{bert} outperforms the best reported result on CoordInv by Conneau et al. (2018).

prefix insertions and lexical substitutions. However, they fall short in capturing rhetorical and semantic corruptions, lexical perturbations and corrupt pronouns. We furthermore find that discourse embedding space encodes subject and object relevant information; however, there is scope for substantial improvement in terms of encoding linguistic properties relevant to discourse coherence. Experiments on coordination inversion further suggest that current models have little capacity at encoding information related to intra-sentential coherence.

We hope this study shall provide further insight into how to frame the task of coherence modeling and improve model performance further. Finally, we make our datasets publicly available for researchers to use to test coherence models.

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*, pages 141–148. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 3(1):1–34.

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

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using entity-based features to model coherence in student essays. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 681–684. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2018. Deep attentive sentence ordering network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4340–4349. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129. Association for Computational Linguistics.

Youmna Farag and Helen Yannakoudakis. 2019. Multi-task learning for coherence modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 629–639, Florence, Italy. 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 *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949. Dublin City University and Association for Computational Linguistics.

Katja Filippova and Michael Strube. 2007. Extending the entity-grid coherence model to semantically related entities. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 139–142, Saarbrücken, Germany. DFKI GmbH.

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

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103. Association for Computational Linguistics.

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

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138.

Shafiq Joty, Muhammad Tasnim Mohiuddin, 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. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223. Association for Computational Linguistics.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 1085–1090, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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. Association for Computational Linguistics.

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*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Lajanugen Logeswaran, Honglak Lee, and Dragomir R. Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In *AAAI*, pages 5285–5292. AAAI Press.

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

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text Interdisciplinary Journal for the Study of Discourse*, pages 243–281.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2262–2272, Hong Kong, China. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael 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 *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

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.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 950–961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

Dat Tien 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.

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*