Evaluating Document Coherence Modeling

Aili Shen[♣], Meladel Mistica[♣], Bahar Salehi[♣],
Hang Li[♢], Timothy Baldwin[♣], Jianzhong Qi[♣]

[♣] The University of Melbourne, Australia
[♢] AI Lab at ByteDance, China
{aili.shen, misticam, tbaldwin, jianzhong.qi}@unimelb.edu.au
baharsalehi@gmail.com, lihang.lh@bytedance.com

Abstract

While pretrained language models (LMs) have driven impressive gains over morpho-syntactic and semantic tasks, their ability to model discourse and pragmatic phenomena is less clear. As a step towards a better understanding of their discourse modeling capabilities, we propose a sentence intrusion detection task. We examine the performance of a broad range of pretrained LMs on this detection task for English. Lacking a dataset for the task, we introduce INSteD, a novel **in**truder **s**en**te**nce **d**etection dataset, containing 170,000+ documents constructed from English Wikipedia and CNN news articles. Our experiments show that pretrained LMs perform impressively in in-domain evaluation, but experience a substantial drop in the cross-domain setting, indicating limited generalization capacity. Further results over a novel linguistic probe dataset show that there is substantial room for improvement, especially in the cross-domain setting.

1 Introduction

Rhetorical relations refer to the transition of one sentence to the next in a span of text (Mann and Thompson, 1988; Asher and Lascarides, 2003). They are important as a discourse device that contributes to the overall coherence, understanding, and flow of the text. These relations span a tremendous breadth of types, including *contrast*, *elaboration*, *narration*, and *justification*. These connections allow us to communicate cooperatively in understanding one another (Grice, 2002; Wilson and Sperber, 2004). The ability to understand such coherence (and conversely detect incoherence) is potentially beneficial for downstream tasks, such as storytelling (Fan et al., 2019; Hu et al., 2020b), recipe generation (Chandu et al., 2019), document-level text generation (Park and Kim, 2015; Holtzman et al., 2018), and essay scoring (Tay et al., 2018; Li et al., 2018).

However, there is little work on document coherence understanding, especially examining the capacity of pretrained language models (LMs) to model the coherence of longer documents. To address this gap, we examine the capacity of pretrained language models to capture document coherence, focused around two research questions: (1) do models truly capture the intrinsic properties of document coherence? and (2) what types of document incoherence can/can't these models detect?

We propose the sentence intrusion detection task: (1) to determine whether a document contains an intruder sentence (coarse-grained level); and (2) to identify the span of any intruder sentence (fine-grained level). We restrict the scope of the intruder text to a single sentence, noting that in practice, the incoherent text could span multiple sentences, or alternatively be sub-sentential.

Existing datasets in document coherence measurement (Chen et al., 2019; Clercq et al., 2014; Lai and Tetreault, 2018; Mim et al., 2019; Pitler and Nenkova, 2008; Tien Nguyen and Joty, 2017) are unsuitable for our task: They are either prohibitively small, or do not specify the span of incoherent text. For example, in the dataset of Lai and Tetreault (2018), each document is assigned a coherence score, but the span of incoherent text is not specified. There is thus a need for a large-scale dataset which includes annotation of the position of intruder text. Identifying the span of incoherent text can benefit tasks where explainability and immediate feedback are important, such as essay scoring (Tay et al., 2018; Li et al., 2018).

In this work, we introduce a dataset consisting of English documents from two domains: Wikipedia articles (106K) and CNN news articles

Transactions of the Association for Computational Linguistics, vol. 9, pp. 621–640, 2021. https://doi.org/10.1162/tacl_a_00388
Action Editor: Noah Smith. Submission batch: 11/2020; Revision batch: 1/2021; Published 7/2021.

(1) Mark Ferguson (born 21 May 1990) is an Irish handballer, currently playing in Dublin, Ireland.
(2) **It is a twelve time Asian Champion, the tournament has been won by any other nation only twice.**
(3) Previously playing for his university team ITB, in the 2013/14 League his club Lughnasa HC came 3rd ...
(4) Mark has been involved in the National Team from 2011 and played in Ireland's first ever European qualifiers ...
(5) The following year Ireland took part in the 2016 Men's European Championship Qualification ...

Figure 1: An excerpt of an incoherent document, with the ''intruder'' sentence indicated in **bold**.

(72K). This dataset fills a gap in research pertaining to document coherence: Our dataset is large in scale, includes both coherent and incoherent documents, and has mark-up of the position of any intruder sentence. Figure 1 is an example document with an intruder sentence. Here, the highlighted sentence reads as though it should be an elaboration of the previous sentence, but clearly exhibits an abrupt change of topic and the pronoun *it* cannot be readily resolved.

This paper makes the following contributions: (1) we propose the sentence intrusion detection task, and examine how pretrained LMs perform over the task and hence at document coherence understanding; (2) we construct a large-scale dataset from two domains—Wikipedia and CNN news articles—that consists of coherent and incoherent documents, and is accompanied with the positions of intruder sentences, to evaluate in both in-domain and cross-domain settings; (3) we examine the behavior of models and humans, to better understand the ability of models to model the intrinsic properties of document coherence; and (4) we further hand-craft adversarial test instances across a variety of linguistic phenomena to better understand the types of incoherence that a given model can detect.

2 Related Work

We first review tasks relevant to our proposed task, then describe existing datasets used in coherence measurement, and finally discuss work on dataset artefacts and linguistic probes.

2.1 Document Coherence Measurement

Coherence measurement has been studied across various tasks, such as the document discrimination task (Barzilay and Lapata, 2005; Elsner et al., 2007; Barzilay and Lapata, 2008; Elsner and Charniak, 2011; Li and Jurafsky, 2017; Putra and Tokunaga, 2017), sentence insertion (Elsner and Charniak, 2011; Putra and Tokunaga, 2017; Xu et al., 2019), paragraph reconstruction (Lapata, 2003; Elsner et al., 2007; Li and Jurafsky, 2017; Xu et al., 2019; Prabhumoye et al., 2020), summary coherence rating (Barzilay and Lapata 2005; Pitler et al., 2010; Guinaudeau and Strube, 2013; Tien Nguyen and Joty, 2017), readability assessment (Guinaudeau and Strube, 2013; Mesgar and Strube, 2016, 2018), and essay scoring (Mesgar and Strube, 2018; Somasundaran et al., 2014; Tay et al., 2018). These tasks differ from our task of intruder sentence detection as follows. First, the document discrimination task assigns coherence scores to a document and its sentence-permuted versions, where the original document is considered to be well-written and coherent and permuted versions incoherent. Incoherence is introduced by shuffling sentences, while our intruder sentences are selected from a second document, and there is only ever a single intruder sentence per document. Second, sentence insertion aims to find the correct position to insert a removed sentence back into a document. Paragraph reconstruction aims to recover the original sentence order of a shuffled paragraph given its first sentence. These two tasks do not consider sentences from outside of the document of interest. Third, the aforementioned three tasks are artificial, and have very limited utility in terms of real-world tasks, while our task can provide direct benefit in applications such as essay scoring, in identifying incoherent (intruder) sentences as a means of providing user feedback and explainability of essay scores. Lastly, in summary coherence rating, readability assessment, and essay scoring, coherence is just one dimension of the overall document quality measurement.

Various methods have been proposed to capture local and global coherence, while our work aims to examine the performance of existing pretrained LMs in document coherence understanding. To assess local coherence, traditional studies have used entity matrices, for example, to represent entity transitions across sentences (Barzilay and

Lapata, 2005, 2008). Guinaudeau and Strube (2013) and Mesgar and Strube (2016) use a graph to model entity transition sequences. Sentences in a document are represented by nodes in the graph, and two nodes are connected if they share the same or similar entities. Neural models have also been proposed (Ji and Smith, 2017; Li and Jurafsky, 2017; Li et al., 2018; Mesgar and Strube, 2018; Mim et al., 2019; Tien Nguyen and Joty, 2017). For example, Tay et al. (2018) capture local coherence by computing the similarity of the output of two LSTMs (Hochreiter and Schmid-huber, 1997), which they concatenate with essay representations to score essays. Li et al. (2018) use multi-headed self-attention to capture long distance relationships between words, which are passed to an LSTM layer to estimate essay coherence scores. Xu et al. (2019) use the average of local coherence scores between consecutive pairs of sentences as the document coherence score.

Another relevant task is disfluency detection in spontaneous speech transcription (Johnson and Charniak, 2004; Jamshid Lou et al., 2018). This task detects the reparandum and repair in spontaneous speech transcriptions to make the text fluent by replacing the reparandum with the repair. Also relevant is language identification in code-switched text (Adouane et al., 2018a,b; Mave et al., 2018; Yirmibeşoğlu and Eryiğit, 2018), where disfluency is defined at the language level (e.g., for a monolingual speaker). Lau et al. (2015) and Warstadt et al. (2019) predict sentence-level acceptability (how natural a sentence is). However, none of tasks are designed to measure document coherence, although sentence-level phenomena can certainly impact on document coherence.

2.2 Document Coherence Datasets

There exist a number of datasets targeted at discourse understanding. For example, Alikhani et al. (2019) construct a multi-modal dataset for understanding discourse relations between text and imagery, such as elaboration and exemplification. In contrast, we focus on discourse relations in a document at the inter-sentential level. The Penn Discourse Treebank (Miltsakaki et al., 2004; Prasad et al., 2008) is a corpus of coherent documents with annotations of discourse connectives and their arguments, noting that inter-sentential discourse relations are not always lexically marked (Webber, 2009).

The most relevant work to ours is the discourse coherence dataset of Chen et al. (2019), which was proposed to evaluate the capabilities of pretrained LMs in capturing discourse context. This dataset contains documents (18K Wikipedia articles and 10K documents from the Ubuntu IRC channel) with fixed sentence length, and labels documents only in terms of whether they are incoherent, without considering the position of the incoherent sentence. In contrast, our dataset: (1) provides more fine-grained information (i.e., the sentence position); (2) is larger in scale (over 170K documents); (3) contains documents of varying length; (4) incorporates adversarial filtering to reduce dataset artefacts (see Section 3); and (5) is accompanied with human annotation over the Wikipedia subset, allowing us to understand behavior patterns of machines and humans.

2.3 Dataset Artefacts

Also relevant to this research is work on removing artefacts in datasets (Zellers et al., 2019; McCoy et al., 2019; Zellers et al., 2018). For example, based on analysis of the SWAG dataset (Zellers et al., 2018), Zellers et al. (2019) find artefacts such as stylistic biases, which correlate with the document labeling and mean that naive models are able to achieve abnormally high results. Similarly, McCoy et al. (2019) examine artefacts in an NLI dataset, and find that naive heuristics that are not directly related to the task can perform remarkably well. We incorporate the findings of such work in the construction of our dataset.

2.4 Linguistic Probes

Adversarial training has been used to craft adversarial examples to obtain more robust models, either by manipulating model parameters (white-box attacks) or minimally editing text at the character/word/phrase level (black-box attacks). For example, Papernot et al. (2018) provide a reference library of adversarial example construction techniques and adversarial training methods.

As we aim to understand the linguistic properties that each model has captured, we focus on black-box attacks (Sato et al., 2018; Cheng et al., 2020; Liang et al., 2018; Yang et al., 2020; Samanta and Mehta, 2017). For example, Samanta and Mehta (2017) construct adversarial examples for sentiment classification and gender detection by deleting, replacing, or inserting words in the text. For a comprehensive review of such studies, see Belinkov and Glass (2019).

There is also a rich literature on exploring what kinds of linguistic phenomena a model has learned (Hu et al., 2020a; Hewitt and Liang, 2019; Hewitt and Manning, 2019; Chen et al., 2019; McCoy et al., 2019; Conneau et al., 2018; Gulordava et al., 2018; Peters et al., 2018; Tang et al., 2018; Blevins et al., 2018; Wilcox et al., 2018; Kuncoro et al., 2018; Tran et al., 2018; Belinkov et al., 2017). The basic idea is to use learned representations to predict linguistic properties of interest. Example linguistic properties are subject–verb agreement or syntactic structure, while representations can be word or sentence embeddings. For example, Marvin and Linzen (2018) construct minimal sentence pairs, consisting of a grammatical and ungrammtical sentence, to explore the capacity of LMs in capturing phenomena such as subject–verb agreement, reflexive anaphora, and negative polarity items. In our work, we hand-construct intruder sentences which result in incoherent documents, based on a broad range of linguistic phenomena.

3 Dataset Construction

3.1 Dataset Desiderata

To construct a large-scale, low-noise dataset that truly tests the ability of systems to detect intruder sentences, we posit five desiderata:

1. **Multiple sources:** The dataset should not be too homogeneous in terms of genre or domain, and should ideally test the ability of models to generalize across domain.

2. **Defences against hacking:** Human annotators and machines should not be able to hack the task and reverse-engineer the labels by sourcing the original documents.

3. **Free of artefacts:** The dataset should be free of artefacts, that allow naive heuristics to perform well.

4. **Topic consistency:** The intruder sentence, which is used to replace a sentence from a coherent document to obtain an incoherent document, should be relevant to the topic of the document, to focus the task on coherence and not simple topic detection.

5. **KB-free:** Our goal is *NOT* to construct a fact-checking dataset; the intruder sentence should be determinable based on the content of the document, without reliance on external knowledge bases or fact-checking.

3.2 Data Sources

We construct a dataset from two sources—Wikipedia and CNN—which differ in style and genre, satisfying the first desideratum. Similar to WikiQA (Yang et al., 2015) and HotpotQA (Yang et al., 2018), we represent a Wikipedia document by its summary section (i.e., the opening paragraph), constraining the length to be between 3 and 8 sentences. For CNN, we adopt the dataset of Hermann et al. (2015) and Nallapati et al. (2016), which consists of over 100,000 news articles. To obtain documents with sentence length similar to those from Wikipedia, we randomly select the first 3–8 sentences from each article.

To defend against dataset hacks[1] that could expose the labels of the test data (desideratum 2), the Wikipedia test set is randomly sampled from 37 historical dumps of Wikipedia, where the selected article has a cosine similarity less than the historical average of 0.72 with its online version.[2] For the training set, we remove this requirement and randomly select articles from different Wikipedia dumps, namely, the articles in the training set might be the same as their current online version. For CNN, we impose no such limitations.

3.3 Generating Candidate Positive Samples

We consider the original documents to be coherent. We construct incoherent documents from half of our sampled documents as follows (satisfying desiderata 3–5):

1. Given a document D, use bigram hashing and TF-IDF matching (Chen et al., 2017) to retrieve the top-10 most similar documents from a collection of documents from the same domain, where D is the query text. Let the set of retrieved documents be $\mathcal{R}_D$.

2. Randomly choose a non-opening sentence S from document D, to be replaced by a sentence candidate generated later. We do

[1] Deliberate or otherwise, e.g., via pre-training on the same version of Wikipedia our dataset was constructed over.

[2] This threshold was determined by calculating the average TF-IDF-weighted similarity of the summary section for documents in all 37 dumps with their current online versions.

not replace the opening sentence as it is needed to establish document context.

3. For each document $D' \in \mathcal{R}_D$, randomly select one non-opening sentence $S' \in D'$ as an intruder sentence candidate.

4. Calculate the TF-IDF-weighted cosine similarity between sentence S and each candidate S'. Remove any candidates with similarity scores ≥ 0.6, to attempt to generate a KB-free incoherence.

5. Replace sentence S with each low-similarity candidate S', and use a fine-tuned XLNet-Large model (Yang et al., 2019) to check whether it is easy for XLNet-Large to detect (see Section 5). For documents with both easy and difficult sentence candidates, we randomly sample from the difficult sentence candidates; otherwise, we randomly choose from all the sentence candidates.

The decision to filter out sentence candidates with similarity ≥ 0.6 was based on the observation that more similar sentences often led to the need for world knowledge to identify the intruder sentence (violating the fifth desideratum). For example, given *It is the* ***second*** *novel in the first of three trilogies about Bernard Samson, ...*, a candidate intruder sentence candidate with high similarity is *It is the* ***first*** *novel in the first of three trilogies about Bernard Samson ....*

We also trialed other ways of generating incoherent samples, such as using sentence S from document D as the query text to retrieve documents, and adopting a 2-hop process to retrieve relevant documents. We found that these methods resulted in documents that can be identified by the pretrained models easily.

4 Dataset Analysis

4.1 Statistics of the Dataset

The process described in Section 3 resulted in 106,352 Wikipedia documents and 72,670 CNN documents, at an average sentence length of 5 in both cases (see Table 1). The percentages of positive samples (46% and 49%, respectively) are slightly less than 50% due to our data generation constraints (detailed in Section 3.3), which can lead to no candidate intruder sentence S' being generated for original sentence S. We set aside 8% of Wikipedia (which we manually tag, as detailed in Section 4.5) and 20% of CNN for testing.

Source	#docs	avg. #sents	avg. #tokens
Wikipedia	106,352 (46%)	5±1	126±24
CNN	72,670 (49%)	5±1	134±32

Table 1: Dataset statistics for INSteD. Numbers in parentheses are percentages of incoherent documents.

4.2 Types of Incoherence

To better understand the different types of issues resulting from our automatic method, we sampled 100 (synthesized) incoherent documents from Wikipedia and manually classified the causes of incoherence according to three overlapping categories (ranked in terms of expected ease of detection): (1) information structure inconsistency (a break in information flow); (2) logical inconsistency (a logically inconsistent world state is generated, such as someone attending school before they were born); and (3) factual inconsistency (where the intruder sentence is factually incorrect). See Table 2 for a breakdown across the categories, noting that a single document can be incoherent across multiple categories. Information structure inconsistency is the most common form of incoherence, followed by factual inconsistency. The 35% of documents with factual inconsistencies break down into 8% (overall) that have other types of incoherence, and 27% that only have a factual inconsistency. This is an issue for the fifth desideratum for our dataset (see Section 3.1), motivating the need for manual checking of the dataset to determine how readily the intruder sentence can be detected.[3]

4.3 Evaluation Metrics

We base evaluation of intruder sentence detection at both the document and sentence levels:

- **document level**: *Does the document contain an intruder sentence?* This is measured based on classification accuracy (Acc), noting that the dataset is relatively balanced at the document level (see Table 1). A prediction

[3] We keep these documents in the dataset, as it is beyond the scope of this work to filter these documents out.

Incoherence	Example	%
Information structure inconsistency	*He is currently the senior pastor at Sovereign Grace Church of Louisville.* ***The Church is led by Senior Pastor Ray Johnston, Senior Pastor Curt Harlow and Senior Pastor Andrew McCourt, and Senior Pastor Lincoln Brewster.*** *Under Mahaney's leadership, Sovereign Grace Church of Louisville is a member of Sovereign Grace Churches.*	58
Logical inconsistency	*Michael David, born September 22, 1954, is an American-born American painter.* ***From 1947–1949 he attended the Otis Art Institute, from 1947 to 1950 he also attended the Art Center College of Design in Los Angeles, and in 1950 the Chouinard Art Institute.***	26
Factual inconsistency	*The Newport Tower has 37 floors.* ***It is located on the beachfront on the east side of Collins Avenue between 68th and 69th Streets.*** *The building was developed by Melvin Simon & Associates in 1990.*	35

Table 2: Types of document incoherence in Wikipedia. Text in **bold** indicates the intruder sentence.

is ''correct'' if at least one sentence/none of the sentences is predicted to be an intruder.

- **sentence level**: *Is a given (non-opening) sentence an intruder sentence?* This is measured based on F_1, noting that most (roughly 88%) sentences are non-intruder sentences.

4.4 Testing for Dataset Artefacts

To test for artefacts, we use XLNet-Large (Yang et al., 2019) to predict whether each non-opening sentence is an intruder sentence, *in complete isolation of its containing document* (i.e., as a standalone sentence classification task). We compare the performance of XLNet-Large with a majority-class baseline (''Majority-class'') that predicts all sentences to be non-intruder sentences (i.e., from the original document), where XLNet-Large is fine-tuned over the Wikipedia/CNN training set, and tested over the corresponding test set.

For Wikipedia, XLNet-Large obtains an Acc of 55.4% (vs. 55.1% for Majority-class) and F_1 of 3.4% (vs. 0.0% for Majority-class). For CNN, the results are 50.8% and 1.2%, respectively (vs. 51.0% and 0.0% resp. for Majority-class). These results suggest that the dataset does not contain obvious artefacts, at least for XLNet-Large. We also experiment with a TF-IDF weighted bag-of-words logistic regression model, achieving slightly worse results than XLNet-Large (Acc = 55.1%, F_1 = 0.05% for Wikipedia, and Acc = 50.6%, F_1 = 0.3% for CNN).[4]

4.5 Human Verification

We performed crowdsourcing via Amazon Mechanical Turk over the Wikipedia test data to examine how humans perform over this task. Each Human Intelligence Task (HIT) contained 5 documents and was assigned to 5 workers. For each document, the task was to identify a single sentence that ''creates an incoherence or break in the content flow'', or in the case of no such sentence, ''None of the above'', indicating a coherent document. In the task instructions, workers were informed that there is at most one intruder sentence per document, and were not able to select the opening sentence. Among the 5 documents for each HIT, there was one incoherent document from the training set, which was pre-identified as being easily detectable by an author of the paper, and acts as a quality control item. We include documents where at least 3 humans assign the same label as our test dataset (90.3% of the Wikipedia test dataset), where all the results are reported over these documents, if not specified.[5] Payment was calibrated to be above Australian minimum wage.

Figure 2 shows the distribution of instances where different numbers of workers produced the correct answer (the red bar). For example, for 6.2% of instances, 2 of 5 workers annotated correctly. The blue bars indicate the proportion of incoherent documents where the intruder sentence was correctly detected by the given number of annotators (e.g., for 9.3% of incoherent documents, only 2 of 5 workers were able to identify the intruder sentence correctly). Humans tend to agree with each other over coherent documents, as indicated by the increasing percentages for

[4] For RoBERTa-Large (Section 6.1), there were also no obvious artefacts observed in the standalone sentence setting: Acc = 55.7% and F_1 = 5.3% over Wikipedia, and Acc = 51.3% and F_1 = 4.3% over CNN.

[5] Different people may have different thresholds in considering a document to be incoherent, but this is beyond the scope of our work.

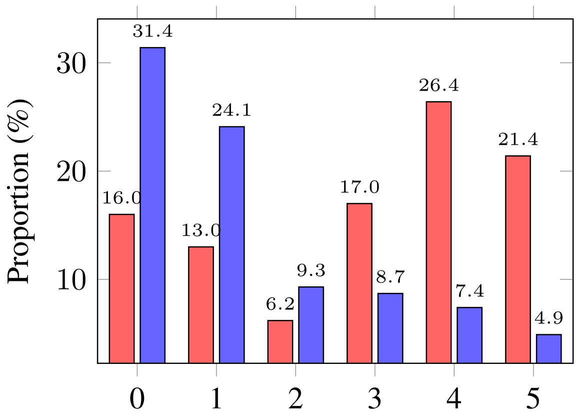

Figure 2: Distribution of instances where different numbers of humans produce correct answers. Note that the red bars indicate distributions over all documents and the blue bars indicate distributions over incoherent documents.

red bars but decreasing percentages for blue bars across the x-axis. Intruder sentences in incoherent documents, however, are harder to detect. One possible explanation is that the identification of intruder sentences requires fact-checking, which workers were instructed not to do (and base their judgment only on the information in the provided document); another reason is that intruder sentences disrupt local incoherence with neighboring sentences, creating confusion as to which is the intruder sentence (with many of the sentence-level mis-annotations being off-by-one errors).

5 Models

We model intruder sentence detection as a binary classification task: Each non-opening sentence in a document is concatenated with the document, and a model is asked to predict whether the sentence is an intruder sentence to the document.

Our focus is on the task, dataset, and how existing models perform at document coherence prediction rather than modeling novelty, and we thus experiment with pre-existing pre-trained models. The models are as follows, each of which is fed into an MLP layer with a softmax output.

BoW: Average the word embeddings for the combined document (sentence + sequence of sentences in the document), based on pretrained 300D GloVe embeddings trained on a 840B-token corpus (Pennington et al., 2014).

Bi-LSTM: Feed the sequence of words in the combined document into a single-layer 512D Bi-LSTM with average-pooling; word embeddings are initialized as with BoW.

InferSent: Generate representations for the sentence and document with InferSent (Conneau et al., 2017), and concatenate the two; InferSent is based on a Bi-LSTM with a max-pooling layer, trained on SNLI (Bowman et al., 2015).

Skip-Thought: Generate representations for the sentence and document with Skip-Thought (Kiros et al., 2015), and concatenate the two; Skip-Thought is an encoder–decoder model where the encoder extracts generic sentence embeddings and the decoder reconstructs surrounding sentences of the encoded sentence.

BERT: Generate representations for the concatenated sentence and document with BERT (Devlin et al., 2019), which was pretrained on the tasks of masked language modeling and next sentence prediction over Wikipedia and BooksCorpus (Zhu et al., 2015); we experiment with both BERT-Large and BERT-Base (the cased versions).

RoBERTa: Generate representations for the concatenated sentence and document with RoBERTa (Liu et al., 2019), which was pretrained on the task of masked language modeling (dynamically masking) and each input consisting of continuous sentences from the same document or multiple documents (providing broader context) over Cc-news, OpenWebTextCorpus, and STORIES (Trinh and Le, 2018), in addition to the same data BERT was pretrained on; we experiment with both RoBERTa-Large and RoBERTa-Base.

ALBERT: Generate representations for the concatenated sentence and document with ALBERT (Lan et al., 2020), which was pretrained over the same dataset as BERT but replaces the next sentence prediction objective with a sentence-order prediction objective, to model document coherence; we experiment with both ALBERT-Large and ALBERT-xxLarge.

XLNet: Generate representations for the concatenated sentence and document with XLNet (Yang et al., 2019), which was pretrained using a permutation language modeling objective over datasets including Wikipedia, BooksCorpus, Giga5 (Parker et al., 2011), ClueWeb 2012-B (Callan et al., 2009), and Common Crawl; we experiment with both XLNet-Largeand XLNet-Base (the cased versions). Although XLNet-Large

	→ Wikipedia				→ CNN			
	Wiki→Wiki		CNN→Wiki		CNN→CNN		Wiki→CNN	
	Acc (%)	F_1 (%)	Acc (%)	F_1 (%)	Acc (%)	F_1 (%)	Acc (%)	F_1 (%)
Majority-class	57.3	0.0	57.3	0.0	50.6	0.0	50.6	0.0
BoW	57.3	0.0	57.3	0.0	50.6	0.0	50.6	0.0
Bi-LSTM	56.2	12.7	57.3	0.0	51.7	25.1	50.2	3.0
InferSent	57.3	0.0	57.3	0.0	50.6	0.0	50.6	0.0
Skip-Thought	57.3	0.0	57.3	0.0	50.6	0.0	50.6	0.0
BERT-Base	65.3	35.7	61.2	21.1	80.8	71.6	57.0	23.5
BERT-Large	67.0	39.6	64.0	29.1	82.4	74.8	61.5	35.9
XLNet-Base	67.8	45.0	62.2	22.4	91.2	86.6	64.0	43.3
XLNet-Large	72.9	55.4	62.8	22.2	**96.9**	95.0	80.7	73.8
RoBERTa-Base	69.5	47.0	63.2	26.1	92.5	88.8	77.6	68.1
RoBERTa-Large	76.1	59.8	63.7	24.6	96.0	94.5	88.3	83.5
ALBERT-Large	70.7	49.6	63.8	24.9	93.4	90.8	72.6	61.5
ALBERT-xxLarge	**81.7**	**71.5**	**66.6**	**33.2**	**96.9**	**95.9**	**89.1**	**86.7**
ALBERT-xxLarge-freeze	57.3	0.0	N/A	N/A	50.6	0.3	N/A	N/A
Human	66.6	35.9	66.6	35.9	74.0		57.8	

Table 3: Experimental results over Wikipedia and CNN, in both in-domain and cross-domain settings. Acc is at the document level and F_1 is at the sentence level.

is used in removing data artefacts when selecting the intruder sentences, our experiments suggest that the comparative results across models (with or without artefact filtering) are robust.

6 Experiments

6.1 Preliminary Results

In our first experiments, we train the various models across both Wikipedia and CNN, and evaluate them in-domain and cross-domain. We are particularly interested in the cross-domain setting, to test the true ability of the model to detect document incoherence, as distinct from overfitting to domain-specific idiosyncrasies. It is also worth mentioning that BERT, RoBERTa, ALBERT, and XLNet are pretrained on multi-sentence Wikipedia data, and have potentially memorised sentence pairs, making in-domain experiments problematic for Wikipedia in particular. Also of concern in applying models to the automatically generated data is that it is entirely possible that an intruder sentence is undetectable to a human, because no incoherence results from the sentence substitution (bearing in mind that only 58% of documents in Table 2 contained information structure inconsistencies).

From Table 3, we can see that the simpler models (BoW, Bi-LSTM, InferSent, and Skip-Thought) perform only at the level of Majority-class at the document level, for both Wikipedia and CNN. At the sentence level (F_1), once again the models perform largely at the level of Majority-class (F_1 = 0.0), other than Bi-LSTM in-domain for Wikipedia and CNN. In the final row of the table, we also see that humans are much better at detecting whether documents are incoherent (at the document level) than identifying the position of intruder sentences (at the sentence level), and that in general, human performance is low. This is likely the result of the fact that there are only 58% of documents in Table 2 containing information structure inconsistencies. We only conducted crowdsourcing over Wikipedia due to budget limitations and the fact that the CNN documents are available online, making dataset hacks possible.[6]

Among the pretrained LMs, ALBERT-xxLarge achieves the best performance over Wikipedia and CNN, at both the document and sentence levels. Looking closer at the Wikipedia results, we find that BERT-Large achieves a higher precision than XLNet-Large (71.0% vs. 60.3%), while XLNet-Large achieves a higher recall (51.3% vs. 27.4%). ALBERT-xxLarge achieves a precision higher than BERT-Large (79.7%) and a recall higher than XLNet-Large (64.9%), leading to the overall

[6] To have a general idea about the difficulty of the CNN dataset, one of the authors annotated 100 documents (50 coherent and 50 incoherent documents), randomly sampled from the test set.

best performance. Over CNN, ALBERT-xxLarge, RoBERTa-Large, and XLNet-Large achieve high precision and recall (roughly 93.0% to 97%).[7] The competitive results for ALBERT-xxLarge over Wikipedia and CNN result from the pretraining strategies, especially the sentence-order prediction loss capturing document coherence in isolation, different from next sentence prediction loss which conflates topic prediction and coherence prediction in a lower-difficulty single task. The performance gap for ALBERT, RoBERTa, and XLNet between the base and large models are bigger than that of BERT, suggesting that they benefit from greater model capacity.[8]

We also examine how pretrained LMs perform with only the classifier parameters being updated during training. Here, we focus on exclusively on ALBERT-xxLarge, given its superiority. As shown in Figure 3, the pretrained LM ALBERT-xxLarge is unable to different coherent documents from incoherent ones, resulting into random guess, although it considers document coherence during pretraining. This indicates the necessity of fine-tuning LMs for document coherent understanding.

Looking to the cross-domain results, again, ALBERT-xxLarge achieves the best performance over both Wikipedia and CNN. The lower results for RoBERTa-Large and XLNet-Large over Wikipedia may be due to both RoBERTa and XLNet being pretrained over newswire documents, and fine-tuning over CNN reducing the capacity of the model to generalize. ALBERT and BERT do not suffer from this as they are not pretrained over newswire documents. The substantial drop between the in- and cross-domain settings for ALBERT, RoBERTa, XLNet, and BERT indicates that the models have limited capacity to learn a generalized representation of document coherence, in addition to the style differences between Wikipedia and CNN.

	Wiki→Wiki	Ubuntu→Wiki
Majority-class	50.0	50.0
ALBERT-xxLarge	96.8	53.1
Human	98.0	98.0
	Ubuntu→Ubuntu	Wiki→Ubuntu
Majority-class	50.0	50.0
ALBERT-xxLarge	58.1	58.7
Human	74.0	74.0

Table 4: Acc for the dataset of Chen et al. (2019).

6.2 Results over the Existing Dataset

We also examine how ALBERT-xxLarge performs over the coarse-grained dataset of Chen et al. (2019), where 50 documents from each domain were annotated by a native English speaker. Performance is measured at the document level only, as the dataset does not include indication of which sentence is the intruder sentence. As shown in Table 4, ALBERT-xxLarge achieves an Acc of 96.8% over the Wikipedia subset, demonstrating that our Wikipedia dataset is more challenging (Acc of 81.7%) and also underlining the utility of adversarial filtering in dataset construction. Given the considerably lower results, one could conclude that Ubuntu is a good source for a dataset. However, when one of the authors attempted to perform the task manually, they found the document-level task to be extremely difficult as it relied heavily on expert knowledge of Ubuntu packages, much more so than document coherence understanding.

In the cross-domain setting, there is a substantial drop over the Wikipedia dataset, which can be explained by ALBERT-Large failing to generate a representation of document coherence from the Ubuntu dataset, due to the high dependence on domain knowledge as described above, resulting in near-random results. The cross-domain results for ALBERT-xxLarge over Ubuntu are actually marginally higher than the in-domain results but still close to random, suggesting that the in-domain model isn't able to capture either document coherence or domain knowledge, and underlining the relatively minor role of coherence for the Ubuntu dataset.

6.3 Performance on Documents of Different Difficulty Levels

One concern with our preliminary experiments was whether the intruder sentences generate genuine incoherence in the information structure of the documents. We investigate this question

[7] The higher performance for all models/humans over the CNN dataset indicates that it is easier for models/humans to identify the presence of intruder sentences. This is can be explained by the fact that a large proportion of documents include named entities, making it easier to detect the intruder sentences. In addition, the database used to retrieve candidate intruder sentences is smaller compared to that of Wikipedia.

[8] We also performed experiments where the models were allowed to predict the first sentence as the intruder sentence. As expected, model performance drops, e.g., F_1 of XLNet-Large drops from 55.4% to 47.9%, reflecting both the increased complexity of the task and the lack of (at least) one previous sentence to provide document context.

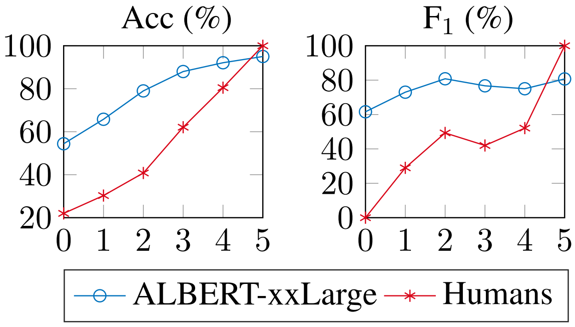

Figure 3: ALBERT-xxLarge vs. humans.

by breaking down the results over the best-performing model (ALBERT-xxLarge) based on the level of agreement between the human annotations and the generated gold-standard, for Wikipedia. The results are in Figure 3, where the x-axis denotes the number of annotators who agree with the gold-standard: for example, ''2'' indicates that 2 of 5 annotators were able to assign the gold-standard labels to the documents.

Our assumption is that the incoherent documents which humans fail to detect are actually not perceptibly incoherent,[9] and that any advantage for the models over humans for documents with low-agreement (with respect to the gold-standard) is actually due to dataset artefacts. At the document level (Acc), there is reasonable correlation between model and human performance (i.e., the model struggles on the same documents as the humans). At the sentence level (F$_1$), there is less discernible difference in model performance over documents of varying human difficulty.

6.4 Analysis over Documents with High Human Agreement

To understand the relationship between human-assigned labels and the gold-standard, we further examine documents where all 5 annotators agree, noting that human-assigned labels can potentially be different from the gold-standard here. Table 5 shows the statistics of humans over these documents, with regard to whether there is an intruder sentence in the documents. Encouragingly, we can see that humans tend to agree more over coherent documents (documents without any intruder sentences) than incoherent documents (documents with an intruder sentence). Examining the 11 original coherent documents which were annotated as incoherent by all annotators, we find out that there is a break in information flow due to references or urls, even though there is no intruder sentence. For documents with an intruder sentence (+ intruder), where humans disagree with the gold-standard (humans perceive the documents as coherent or the position of the intruder sentence to be other than the actual intruder sentence), we find that 98% of the documents are considered to be coherent. We randomly sampled 100 documents from these documents and examined whether the intruder sentence results in a break in information flow. We find that fact-checking is needed to identify the intruder sentence for 93% of the documents.[10]

Humans	# −intruder docs	# +intruder docs
Coherent	1385	177
Incoherent	11	404

Table 5: Statistics over documents where all 5 humans agree, where −intruder/+intruder indicates the documents without/with an intruder sentence.

Table 6 shows the performance over the Wikipedia documents that are annotated consistently by all 5 annotators (from Table 5). Consistent with the results from Table 3, ALBERT-xxLarge achieves the best performance both in- and cross-domain. To understand the different behaviors of humans and ALBERT-xxLarge we analyze documents which only humans got correct, only ALBERT-xxLarge got correct, or neither humans nor ALBERT-xxLarge got correct, as follows:

1. Humans only: 7 incoherent (+intruder) and 73 coherent (−intruder) documents

2. ALBERT-xxLarge only: 181 incoherent (+intruder) (of which we found 97% to require fact-checking[11]) and 9 coherent (−intruder) documents (of which 8 contain urls/references, which confused humans)

[9] Although the intruder sentence may lead to factual errors, the annotators were instructed not to do fact checking.

[10] Here, the high percentage of incoherent documents with factual inconsistencies does not necessarily point to a high percentage of factual inconsistency in the overall dataset, as humans are more likely to agree with the gold-standard for coherent documents.

[11] There are 4 documents that humans identify as incoherent based on the wrong intruder sentence, due to the intruder sentence leading to a misleading factual inconsistency.

	Wiki→Wiki		CNN→Wiki	
	Acc (%)	F_1 (%)	Acc (%)	F_1 (%)
Majority-class	70.6	0.0	70.6	0.0
BERT-Large	76.7	42.0	75.4	36.9
XLNet-Large	79.1	57.0	76.6	35.4
RoBERTa-Large	82.0	59.6	77.3	37.4
ALBERT-xxLarge	**85.9**	**68.8**	**78.8**	**42.9**
Human	79.5	45.4	79.5	45.4

Table 6: Results over documents annotated consistently by all 5 annotators, where annotations can be the same as or different from gold-standard.

3. Neither humans nor models: 223 incoherent (+intruder) (of which 98.2% and 77.1% were predicted to be coherent by humans and ALBERT-xxLarge, respectively, and for the remainder, the wrong intruder sentence was identified) and 2 coherent (−intruder) documents (both of which were poorly organised, confusing allcomers)

Looking over the incoherent documents that require fact-checking, no obvious differences are discernible between the documents that ALBERT-xxLarge predicts correctly and those it misses. Our assumption here is that ALBERT-xxLarge is biased by the pretraining dataset, and that many of the cases where it makes the wrong prediction are attributable to mismatches between the text in our dataset and the Wikipedia version used in pretraining the model.

6.5 Question Revisited

Q1: Do models truly capture the intrinsic properties of document coherence?

A: It is certainly true that models that incorporate a more explicit notion of document coherence into pretraining (e.g., ALBERT) tend to perform better. In addition, larger-context models (RoBERTa) and robust training strategies (XLNet) during pretraining are also beneficial for document coherent understanding. This suggests a tentative yes, but there were equally instances of strong disagreement with human intuitions and model predictions for the better-performing models and evidence to suggest that the models were performing fact-checking at the same time as coherence modeling.

Q2: What types of document incoherence can/can't these models detect?

A: Over incoherent documents resulting from fact inconsistencies, where humans tend to fail, the better-performing models can often make correct predictions; over incoherent documents with information structure or logical inconsistencies which humans can easily detect, ALBERT-Large, RoBERTa-Large, and XLNet-Large achieve an Acc $\geq$ 87%, showing that they can certainly capture information structure and logical inconsistencies to a high degree. That said, the fact that they misclassify clearly coherent documents as incoherent suggests that are in part lacking in their ability to capture document coherence. We thus can conclude that they can reliably identify intruder sentences which result in a break in information structure or logical flow, but are imperfect models of document coherence.

7 Linguistic Probes

To further examine the models, we constructed a language probe dataset.

7.1 Linguistic Probe Dataset Construction

We handcrafted adversarial instances based on a range of linguistic phenomena that generate information structure inconsistencies. In constructing such a dataset, minimal modifications were made to the original sentences, to isolate the effect of the linguistic probe. For each phenomenon, we hand-constructed roughly 100 adversarial instances by modifying intruder sentences in *incoherent* Wikipedia test documents that were manually pre-filtered for ease of detection/lack of confounding effects in the original text. That is, the linguistic probes for the different phenomena were manually added to incoherent test documents, within intruder sentences; our interest here is whether the addition of the linguistic probes makes it easier for the models to detect the incoherence. Importantly, we do not provide any additional training data, meaning there is no supervision signal specific to the phenomena. There are roughly 8×100 instances in total,[12] with the eight phenomena being:

1. gender pronoun flip (*Gender*), converting a pronoun to its opposite gender (e.g., *she* → *he*);

[12] There are 100 instances for each phenomenon except for *Demonstrative*, where there were only 95 instances in the Wikipedia test data with singular demonstratives.

2. animacy downgrade (*Animacy*$\downarrow$), downgrading pronouns and possessive determiners to their inanimate versions (e.g., *she/he/her/him* $\rightarrow$ *it*, and *her/his* $\rightarrow$ *its*);
3. animacy upgrade (*Animacy*$\uparrow$), upgrading pronouns and possessive determiners to their third person version (e.g., *it* $\rightarrow$ *she/he/her/him*, and *its* $\rightarrow$ *her/his*);
4. singular demonstrative flip (*Demonstrative*), converting singular demonstratives to plural ones (e.g., *this* $\rightarrow$ *these* and *that* $\rightarrow$ *those*);
5. conjunction flip (*Conjunction*), converting conjunctions to their opposites (e.g., *but* $\rightarrow$ *and therefore*, *and* $\rightarrow$ *but*, *although* $\rightarrow$ *therefore*, and vice versa);
6. past tense flip (*Past to Future*), converting past to future tense (e.g., *was* $\rightarrow$ *will be* and *led* $\rightarrow$ *will lead*);
7. sentence negation (*Negation*), negating the sentence (e.g., *He has [a] . . . warrant ...* $\rightarrow$ *He doesn't have [a] . . . warrant ...*);
8. number manipulation (*Number*), changing numbers to implausible values (e.g., *He served as Chief Operating Officer . . . from 2002 to 2005* $\rightarrow$ *He served as Chief Operating Officer . . . from 200 BCE to 201 BCE* and *Line 11 has a length of 51.7 km and a total of 18 stations.* $\rightarrow$ *Line 11 has a length of 51.7 m and a total of 1.8 stations.*).

All the probes generate syntactically correct sentences, and the first four generally lead to sentences that are also semantically felicitous, with the incoherence being at the document level. For example, in *He was never convicted and was out on parole within a few years*, if we replace *he* with *she*, the sentence is felicitous, but if the focus entity in the preceding and subsequent sentences is a male, the information flow will be disrupted.

The last four language probes are crafted to explore the capacity of a model to capture commonsense reasoning, in terms of discourse relationships, tense and polarity awareness, and understanding of numbers. For *Conjunction*, we only focus on explicit connectives within a sentence. For *Past to Future*, there can be intra-sentence inconsistency if there are time-specific signals, failing which broader document context is needed to pick up on the tense flip. Similarly for *Negation* and *Number*, the change can lead to inconsistency either intra- or inter-sententially. For example, *He did not appear in more than 400 films between 1914 and 1941* ... is intra-sententially incoherent.

7.2 Experimental Results

Table 7 lists the performance of pretrained LMs at recognising intruder sentences within incoherent documents, with and without the addition of the respective linguistic probes.[13] For a given model, we break down the results across probes into two columns: The first column (''F_1'') shows the sentence-level performance over the original intruder sentence (without the addition of the linguistic probe), and the second column (''$\Delta\mathrm{F}_1$'') shows the absolute difference in performance with the addition of the linguistic probe. Our expectation is that results should improve on average with the inclusion of the linguistic probe (i.e., $\Delta\mathrm{F}_1$ values should be positive), given that we have reinforced the incoherence generated by the intruder sentence.

All models achieve near-perfect results with *Gender* linguistic probes (i.e., the sum of F_1 and $\Delta\mathrm{F}_1$ is close to 100), and are also highly successful at detecting *Animacy* mismatches and *Past to Future* (the top half of Table 7). For the probes in the bottom half of the table, none of the three models except ALBERT-xxLarge performs particularly well, especially for *Demonstrative*. For each linguistic probe, we observe that the pretrained LMs can more easily detect incoherent text with the addition of these lexical/grammatical inconsistencies (except for XLNet-Large and ALBERT-xxLarge over *Demonstrative* and ALBERT-xxLarge over *Conjunction*).

In the cross-domain setting, the overall performance of XLNet-Large$_{\mathrm{CNN}}$ and ALBERT-xxLarge$_{\mathrm{CNN}}$ drops across all linguistic probes, but the absolute gain through the inclusion of the linguistic probe is almost universally larger, suggest that while domain differences hurt the models, they are attuned to the impact of linguistic probes on document coherence and thus learning some more general properties of document (in)coherence. On the other hand, BERT-Large$_{\mathrm{CNN}}$ (over *Gender*, *Animacy*$\downarrow$, and *Animacy*$\uparrow$) and RoBERTa-Large$_{\mathrm{CNN}}$ (*Gender* and *Animacy*$\uparrow$) actually perform better than in-domain. RoBERTa-Large$_{\mathrm{CNN}}$ achieves the best overall

[13] Results for coherent documents are omitted due to space.

	Gender		*Animacy↓*		*Animacy↑*		*Past to Future*	
	F_1	ΔF_1	F_1	ΔF_1	F_1	ΔF_1	F_1	ΔF_1
BERT-Large	26.5	+65.3	26.3	+53.2	33.6	+45.1	35.6	+42.1
XLNet-Large	55.8	+41.6	50.0	+45.2	64.0	+23.5	64.9	+16.9
RoBERTa-Large	64.9	+32.5	50.7	+38.3	59.7	+21.7	69.2	+19.9
ALBERT-xxLarge	74.0	+25.4	71.8	+8.5	81.0	+2.9	79.8	+4.3
BERT-Large$_{\text{CNN}}$	23.9	+70.0	22.2	+60.2	27.6	+51.4	30.6	+14.7
XLNet-Large$_{\text{CNN}}$	13.6	+83.1	10.0	+71.3	8.0	+71.8	23.2	+27.6
RoBERTa-Large$_{\text{CNN}}$	15.4	+82.4	7.9	+64.4	9.8	+73.3	23.4	+40.0
ALBERT-xxLarge$_{\text{CNN}}$	21.6	+72.8	20.2	+51.8	27.6	+33.4	38.0	+30.4
Human	35.8	+53.4	36.6	+45.3	29.8	+53.9	40.9	+34.4

	Conjunction		*Demonstrative*		*Negation*		*Number*	
	F_1	ΔF_1	F_1	ΔF_1	F_1	ΔF_1	F_1	ΔF_1
BERT-Large	51.9	+17.3	34.8	+15.6	34.5	+32.2	32.5	+31.2
XLNet-Large	68.6	+3.6	55.4	0.0	57.7	+8.9	50.7	+11.3
RoBERTa-Large	73.0	+0.7	57.9	0.0	68.4	+10.9	54.2	+20.0
ALBERT-xxLarge	83.5	−1.6	75.2	+1.3	79.5	+2.9	63.9	+10.4
BERT-Large$_{\text{CNN}}$	38.2	−1.4	35.6	−5.7	28.8	+4.2	19.6	+11.7
XLNet-Large$_{\text{CNN}}$	31.0	0.0	14.1	0.0	15.7	+11.8	15.2	+13.1
RoBERTa-Large$_{\text{CNN}}$	33.9	+1.4	17.8	0.0	21.0	+12.4	18.3	+23.6
ALBERT-xxLarge$_{\text{CNN}}$	41.6	+1.3	30.9	0.0	28.1	+19.2	23.0	+16.0
Human	40.5	+8.7	38.0	+1.0	40.4	+36.8	37.3	+24.2

Table 7: Results over language probes in incoherent Wikipedia test documents. BERT-Large$_{\text{CNN}}$, XLNet-Large$_{\text{CNN}}$, RoBERTa-Large$_{\text{CNN}}$, and ALBERT-xxLarge$_{\text{CNN}}$ are trained over CNN, while BERT-Large, XLNet-Large, RoBERTa-Large, and ALBERT-xxLarge are trained over Wikipedia. Here, F_1 is over the original incoherent documents (excluding linguistic probes), and ΔF_1 indicates the absolute performance difference resulting from incorporating linguistic probes.

performance over *Gender*, *Animacy↑*, and *Number* while ALBERT-xxLarge$_{\text{CNN}}$ achieves the best overall performance over *Past to Future*, *Conjunction*, *Demonstrative*, and *Negation*. The reason that the models tend to struggle with *Demonstrative* and *Conjunction* is not immediately clear, and will be explored in future work.

We also conducted human evaluations on this dataset via Amazon Mechanical Turk, based on the same methodology as described in Section 4.5 (without explicit instruction to look out for linguistic artefacts, and with a mixture of coherent and incoherent documents, as per the original annotation task). As detailed in Table 7, humans generally benefit from the inclusion of the linguistic probes. Largely consistent with the results for the models, humans are highly sensitised to the effects of *Gender*, *Animacy*, *Past to Future*, and *Negation*, but largely oblivious to the effects of *Demonstrative* and *Conjunction*. Remarkably, the best models (ALBERT-xxLarge and RoBERTa-Large) perform on par with humans in the in-domain setting, but are generally well below humans in the cross-domain setting.

8 Conclusion

We propose the new task of detecting whether there is an intruder sentence in a document, generated by replacing an original sentence with a similar sentence from a second document. To benchmark model performance over this task, we construct a large-scale dataset consisting of documents from English Wikipedia and CNN news articles. Experimental results show that pretrained LMs that incorporate larger document contexts in pretraining perform remarkably well in-domain, but experience a substantial drop cross-domain. In follow-up analysis based on human annotations, substantial divergences from human intuitions were observed, pointing to limitations in their ability to model document coherence. Further results over a linguistic probe dataset show that pretrained models fail to identify some linguistic characteristics that affect document coherence, suggesting room to improve for them to truly capture document coherence, and motivating the construction of a dataset with intruder text at the intra-sentential level.

References

Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik. 2018a. Improving neural network performance by injecting background knowledge: Detecting code-switching and borrowing in Algerian texts. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 20–28. **DOI:** `https://doi.org/10.18653/v1/W18-3203`

Wafia Adouane, Simon Dobnik, Jean-Philippe Bernardy, and Nasredine Semmar. 2018b. A comparison of character neural language model and bootstrapping for language identification in multilingual noisy texts. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pages 22–31. **DOI:** `https://doi.org/10.18653/v1/W18-1203`

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A corpus of image-text discourse relations. 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 570–575.

Nicholas M. Asher and Alex Lascarides. 2003. *Logics of Conversation*, Studies in Natural Language Processing. Cambridge University Press.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 141–148. **DOI:** `https://doi.org/10.3115/1219840.1219858`

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34. **DOI:** `https://doi.org/10.1162/coli.2008.34.1.1`

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872. **DOI:** `https://doi.org/10.18653/v1/P17-1080`

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72. **DOI:** `https://doi.org/10.1162/tacl_a_00254`

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep RNNs encode soft hierarchical syntax. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19. **DOI:** `https://doi.org/10.18653/v1/P18-2003`

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. **DOI:** `https://doi.org/10.18653/v1/D15-1075`

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. Clueweb09 data set. `https://lemurproject.org/clueweb09/`. Accessed: 15.12.2019.

Khyathi Chandu, Eric Nyberg, and Alan W Black. 2019. Storyboarding of recipes: Grounded contextual generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6040–6046. **DOI:** `https://doi.org/10.18653/v1/P19-1606`

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879. **DOI:** https://doi.org/10.18653/v1/P17-1171

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 649–662. **DOI:** https://doi.org/10.18653/v1/D19-1060

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 3601–3608. **DOI:** https://doi.org/10.1609/aaai.v34i04.5767

Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325. **DOI:** https://doi.org/10.1017/S1351324912000344

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680. **DOI:** https://doi.org/10.18653/v1/D17-1070

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. **DOI:** https://doi.org/10.18653/v1/P18-1198

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.

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.

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.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660.

Herbert Paul Grice. 2002. Logic and conversation. *Foundations of Cognitive Psychology*, 719–732.

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.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205. **DOI:** https://doi.org/10.18653/v1/N18-1108

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015.

Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, pages 1693–1701.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743. **DOI:** https://doi.org/10.18653/v1/D19-1275

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. **DOI:** https://doi.org/10.1162/neco.1997.9.8.1735, **PMID:** 9377276

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1638–1649. **DOI:** https://doi.org/10.18653/v1/P18-1152

Jennifer Hu, Sherry Y. Chen, and Roger P. Levy. 2020a. A closer look at the performance of neural language models on reflexive anaphor licensing. *Proceedings of the Society for Computation in Linguistics*, 3(1):382–392.

Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2020b. What makes a good story? Designing composite rewards for visual storytelling. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7969–7976. **DOI:** https://doi.org/10.1609/aaai.v34i05.6305

Paria Jamshid Lou, Peter Anderson, and Mark Johnson. 2018. Disfluency detection using auto-correlational neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4610–4619. **DOI:** https://doi.org/10.18653/v1/D18-1490

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005.

Mark Johnson and Eugene Charniak. 2004. A TAG-based noisy-channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 33–39. **DOI:** https://doi.org/10.3115/1218955.1218960

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, pages 3294–3302.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436. **DOI:** https://doi.org/10.18653/v1/P18-1132

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. **DOI:** https://doi.org/10.18653/v1/W18-5023

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations*.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting* *of the Association for Computational Linguistics*, pages 545–552. **DOI:** `https://doi.org/10.3115/1075096.1075165`

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1618–1628.

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.

Xia Li, Minping Chen, Jianyun Nie, Zhenxing Liu, Ziheng Feng, and Yingdan Cai. 2018. Coherence-based automated essay scoring using self-attention. In *Proceedings of the 17th China National Conference on Computational Linguistics, CCL 2018, and the 6th International Symposium on Natural Language Processing Based on Naturally Annotated Big Data*, pages 386–397. **DOI:** `https://doi.org/10.1007/978-3-030-01716-3_32`

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4208–4215. **DOI:** `https://doi.org/10.24963/ijcai.2018/585`

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, cs.CL/1907.11692v1.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281. **DOI:** `https://doi.org/10.1515/text.1.1988.8.3.243`

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. **DOI:** `https://doi.org/10.18653/v1/D18-1151`

Deepthi Mave, Suraj Maharjan, and Thamar Solorio. 2018. Language identification and analysis of code-switched social media text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 51–61. **DOI:** `https://doi.org/10.18653/v1/W18-3206`

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448. **DOI:** `https://doi.org/10.18653/v1/P19-1334`

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1414–1423. **DOI:** `https://doi.org/10.18653/v1/N16-1167`

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. **DOI:** `https://doi.org/10.18653/v1/D18-1464`

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn discourse treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation.*

Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. 2019. Unsupervised learning of discourse-aware text representation. In *Proceedings of the 2019 Student Research Workshop*, pages 93–104.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, 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. **DOI:** https://doi.org/10.18653/v1/K16-1028

Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. 2018. Technical report on the CleverHans v2.1.0 adversarial examples library. *CoRR*, cs.LG/1610.00768v6.

Cesc C. Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In *Proceedings of Advances in Neural Information Processing Systems 28*, pages 73–81.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition, linguistic data consortium. *Google Scholar*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543. **DOI:** https://doi.org/10.3115/v1/D14-1162

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509. **DOI:** https://doi.org/10.18653/v1/D18-1179

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554.

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. **DOI:** https://doi.org/10.3115/1613715.1613742

Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2020. Topological sort for sentence ordering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2783–2792. **DOI:** https://doi.org/10.18653/v1/2020.acl-main.248

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Jan Wira Gotama Putra and Takenobu Tokunaga. 2017. Evaluating text coherence based on semantic similarity graph. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pages 76–85.

Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *CoRR*, cs.LG/1707.02812v1.

Motoki Sato, Jun Suzuki, Hiroyuki Shindo, and Yuji Matsumoto. 2018. Interpretable adversarial perturbation in input embedding space for text. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4323–4330. **DOI:** https://doi.org/10.24963/ijcai.2018/601

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

Gongbo Tang, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? A targeted evaluation of neural machine translation architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing, pages 4263–4272. **DOI:** https://doi.org/10.18653/v1/D18-1458

Yi Tay, Minh C. Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5948–5955.

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. **DOI:** https://doi.org/10.18653/v1/P17-1121

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736. **DOI:** https://doi.org/10.18653/v1/D18-1503

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, cs.AI/1806.02847v2.

Alex Warstadt, Amanpreet Singh, and Samuel Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7(0). **DOI:** https://doi.org/10.1162/tacl_a_00290

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 674–682. **DOI:** https://doi.org/10.3115/1690219.1690240

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221. **DOI:** https://doi.org/10.18653/v1/W18-5423

Deirdre Wilson and Dan Sperber. 2004. Relevance theory, *The Handbook of Pragmatics*, Blackwell, pages 607–632.

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.

Puyudi Yang, Jianbo Chen, Cho-Jui Hsieh, Jane-Ling Wang, and Michael I. Jordan. 2020. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. *Journal of Machine Learning Research*, 21:43:1–43:36.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018. **DOI:** https://doi.org/10.18653/v1/D15-1237

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the Thirty-third Conference on Neural Information Processing Systems*, pages 5754–5764.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. **DOI:** https://doi.org/10.18653/v1/D18-1259

Zeynep Yirmibeşoğlu and Gülşen Eryiğit. 2018. Detecting code-switching between Turkish-English language pair. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 110–115. **DOI:** https://doi.org/10.18653/v1/W18-6115

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104. **DOI:** `https://doi.org/10.18653/v1/D18-1009`

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800. **DOI:** `https://doi.org/10.18653/v1/P19-1472`

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 19–27. **DOI:** `https://doi.org/10.1109/ICCV.2015.11`