Assessing the Coherence Modeling Capabilities of Pretrained Transformer-based Language Models

Anonymous ACL submission

Abstract

The task of ordering a shuffled set of sentences into a coherent text is used to evaluate the capacity of a model to understand causal and temporal relations between entities and events. Recent approaches rely on pretrained Transformer-based models, but it remains unknown whether the differences between them, such as size, pretraining data and objectives, affect their coherence modeling capacity. We present a simple architecture for sentence ordering that relies exclusively on pretrained Transformer-based encoder-only models. This allows us to compare the coherence modeling capabilities of the monolingual and multilingual versions of BERT, RoBERTa, and DistilBERT. We show that RoBERTa-based models outperform BERT-based models and are more robust when ordering longer documents with more than 10 sentences. Thus, the intuitive advantage offered by sentence-based objectives such as Next Sentence Prediction used in BERT is effectively compensated by the higher amount and diversity of the training data used in RoBERTa. However, the difference between multilingual versions of BERT and RoBERTa is narrower. This suggests that exposure to different languages partially makes up for the benefits of larger and more diverse training data.

1 Introduction

As an essential element of discourse, a large number of works have focused on studying coherence; cf., e.g., (Wang and Guo, 2014). Textual coherence refers to the relations of meaning between sentences or propositions of a text, allowing it to be logical and semantically consistent. It requires an understanding of entities, events, and the relations between them. It has a wide range of applications in tasks such as multi-document extractive summarization (Barzilay and Elhadad, 2002; Galanis et al., 2012; Nallapati et al., 2017), question answering (Yu et al., 2018; Liu et al., 2018) and text generation (Konstas and Lapata, 2013; Schwartz et al., 2017; Holtzman et al., 2018).

The sentence ordering task (Barzilay and Lapata, 2008) is commonly used to train and evaluate coherence modeling systems. It aims at finding the most coherent permutation of sentences among all possible orders in a paragraph. Early works exploited linguistic features (Elsner and Charniak, 2008; Lapata et al., 2005; Barzilay and Lapata, 2005; Elsner and Charniak, 2011a; Louis and Nenkova, 2012) and the first neural networks-based approaches relied on pointer networks (Gong et al., 2016; Logeswaran et al., 2018; Cui et al., 2018; Yin et al., 2019, 2020; Wang and Wan, 2019; Oh et al., 2019). However, the recent success of transfer learning from Transformer-based language models in a wide range of cross-lingual transfer tasks has pushed the state-of-the-art much further (Kumar et al., 2020; Prabhumoye et al., 2020; Zhu et al., 2021b,a; Cui et al., 2020; Chowdhury et al., 2021).

There is a wide variety of Transformer-based models that are pretrained with different objectives and data, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), or DistilBERT (Sanh et al., 2019). However, the majority of sentence ordering works rely exclusively on BERT, originally trained with the Masked Language Model (MLM) and Next Sentence Prediction (NSP) objectives on English data. It remains unknown whether optimizing for different pretraining objectives, with different data and different languages, affects the coherence modeling capacity of Transformer-based language models. It seems intuitive that the NSP objective present in BERT but not in RoBERTa should be useful for the sentence ordering task. However, it is also possible that pretraining models with larger and more diverse data as done in RoBERTa can compensate for the lack of sentence-based objectives. Moreover, multilingual models that are trained on multiple languages can benefit by being exposed to different data, e.g. sentences of different length and complexity.

We explore these and other aspects of coherence modeling via a clean and simple architecture for sentence ordering that relies only on pretrained Transformer-based models as the encoder. Our architecture comprises of a document encoder that captures the relations between the sentences and generates a representation for each sentence. The model generates a score for each sentence based on those representations, which is used to sort them. We train and evaluate the monolingual and multilingual versions of RoBERTa, BERT and DistilBERT on five sentence ordering datasets (§4.2). Despite its simplicity, the proposed method is competitive and outperforms more complex models (Prabhumoye et al., 2020; Kumar et al., 2020).

We show that, despite the intuitive advantage of the NSP objective used to train BERT-based models, the coherence modeling capabilities of RoBERTa-based models are stronger than those of BERT-based models. Thus, the larger and more diverse training data used in RoBERTa compensates for the lack of sentence-based objectives. The performance difference between BERT-based and RoBERTa-based models is narrower for the multilingual models, suggesting that exposure to different languages, with e.g. sentences of different length and complexity, partially makes up for the benefits of larger and more diverse training data. Distilled models are close in accuracy to the original models while being lighter and much faster to train. Our main contributions are:

- A simple yet competitive Transformer-based architecture for sentence ordering (§3.1).
- A novel data augmentation strategy designed to leverage as much knowledge as possible from the available data while keeping the training procedure tractable (§3.3).
- A thorough comparison of the coherence modeling abilities of different pretrained Transformer-based language models: Analyzing 1) the utility of the NSP pretraining objective by comparing BERT and RoBERTa-based models; 2) the benefit of using multilingual models in monolingual downstream tasks; and 3) the impact of model size in the coherence modeling capabilities of the models (§5).

2 Related work

Early approaches to sentence ordering focused on modeling local coherence using linguistic features (Lapata et al., 2005; Barzilay and Lapata, 2008; Elsner and Charniak, 2011b; Guinaudeau and Strube, 2013). The first neural network-based approaches relied on pointer networks (Vinyals et al., 2015) to retrieve the correct order by pair-wise comparisons of encoded sentences (Gong et al., 2016; Logeswaran et al., 2018; Cui et al., 2018; Yin et al., 2019, 2020). Later works used pointer networks for decoding (Wang and Wan, 2019; Oh et al., 2019), introducing the use of attention mechanisms (Bahdanau et al., 2014).

Recent approaches use ranking or sorting frameworks for this task. RankTxNet (Kumar et al., 2020) uses BERT sentence representations to compute a score for each sentence, and sorts all the scores with a ranking-based loss function. B-TSort (Prabhumoye et al., 2020) predicts the correct constraint between sentence pairs and uses topological sorting to find the final order. Zhu et al. (2021b) use constraint graphs to generate order-enhanced sentence representations. BERT4SO (Zhu et al., 2021a) presents a BERT-base approach that jointly encodes all sentences instead of encoding each sentence separately, and proposes a margin-based listwise ranking loss. BERSON (Cui et al., 2020) introduces a new relational pointer decoder that incorporates the relative ordering information into the pointer network with a BERT-based deep relational module. While most approaches use BERT, the state-of-the-art approach Re-BART (Chowdhury et al., 2021) is a sequence-to-sequence model that formulates sentence ordering as a conditional text generation task using BART (Lewis et al., 2020).

We present a simplified yet competitive version of Zhu et al.'s BERT4SO. We completely remove the document encoder and simplify the input encoding and the loss function, and our results improve those of BERT4SO. Moreover, the simple and clean architecture allows us to experiment with different pretrained Transformer-based models to study whether optimizing for different model sizes, languages and pretraining objectives affects the coherence modeling capacity of the models.

3 Transformer-based Sentence Ordering

The sentence ordering problem aims at finding the most coherent permutation of sentences among all possible orders in a paragraph. Formally, given a set of N sentences $\{S_{o_1}, S_{o_2}, ..., S_{o_N}\}$, with random order $[o_1, o_2, ..., o_N]$, the model aims to recover the correct order $[o_1^*, o_2^*, ..., o_N^*]$. Following existing work (Kumar et al., 2020; Prabhumoye et al., 2020; Zhu et al., 2021b,a), we frame the task as a ranking problem. We train a model to predict a score z_i for each sentence S_i, and to determine the predicted order by sorting all scores from higher (first sentence) to lower (last sentence).

3.1 Model architecture

We present a clean and simple architecture for sentence ordering that relies exclusively on a Transformer-based model, which is used as the encoder (Figure 1). We experiment with both BERT-based and RoBERTa-based models, and in what follows we will refer to them generally as Pretrained Language Models (PLM).

Input encoding. PLMs are trained with a maximum of two sentences as input, and therefore they are not directly applicable for sentence ordering, where the model has to handle multi-sentence documents. To overcome this obstacle, some works encode each sentence separately (Kumar et al., 2020), while others encode sentence pairs (Prabhumoye et al., 2020; Zhu et al., 2021b). Following Zhu et al. (2021a), we concatenate all sentences into a single sequence, separating each sentence with a `[CLS]` token (BERT) or a `<s>` token (RoBERTa). Each input sequence starts with a `[CLS]` token and ends with a `[SEP]` token (BERT) or a `</s>` token (RoBERTa). If the input length exceeds the model capacity (512 tokens) at training time, we randomly remove sentences.[1]

Sentences encoding. After concatenating the tokens of the different sentences, three different embeddings are added up as input to the encoder: token embeddings, segment embeddings and position embeddings. As shown in Figure 1, the alternation of segment embeddings is used to indicate the sentence to which each token belongs. The output of the `[CLS]` token preceding each sentence is used as sentence representation to compute the score.

Score generation. Once generated, the representation of each sentence is fed into a 2-layer Perceptron in order to generate a score, which is then used to order the input sentences from higher score (first sentence) to lower score (last sentence).

3.2 Loss function

We use ListMLE (Xia et al., 2008), a listwise ranking loss that minimizes a likelihood loss function defined on the predicted list and the ground-truth list. ListMLE has been shown to perform better than pointwise or pairwise losses in optimizing sentence ordering methods (Kumar et al., 2020). Let $\boldsymbol{o} = [o_1, o_2, ..., o_{m_n}]$ be the correct order of a document n containing m sentences. Then,

$$\mathcal{L}_{ListMLE}(\boldsymbol{p}_n) = -logP_M(\boldsymbol{o}|\boldsymbol{p}_n) \qquad (1)$$

$$P_M(\boldsymbol{o}|\boldsymbol{p}_n) = \prod_{k=1}^{m_n} \frac{exp(z_{o_k})}{\sum_{i=k}^{m_n} exp(z_{o_i})} \qquad (2)$$

3.3 Data augmentation strategy

The common approach to neural sentence ordering relies mainly on shuffling the sentences in each document and training the model to predict the correct order (Kumar et al., 2020; Prabhumoye et al., 2020; Zhu et al., 2021b,a; Cui et al., 2020; Chowdhury et al., 2021). Since n distinct objects can have $n!$ permutations, we could potentially generate a huge amount of training examples. However, training with all possible permutations would be extremely time-consuming, and thus we propose a novel training strategy that aims at leveraging as much knowledge as possible from the data while keeping the training procedure efficient. We start by generating a random shuffle order for each document, to compose our default training set. Then, at each epoch we train with the default training set, augmented with 1) a percentage of the examples with a different randomly generated shuffle order; and 2) a percentage of documents from which we randomly remove sentences to generate harder examples. By removing sentences, we are removing semantic content, which difficults recovering the correct order. After experimenting with different combinations of percentages, we select the combination that offers the best accuracy over the validation sets: augment with 100% of documents with different shuffle order, and 25% of documents with one randomly-removed sentence.

[1] Removing sentences generates harder examples, because part of the semantic content necessary to recover the correct order may be potentially removed.

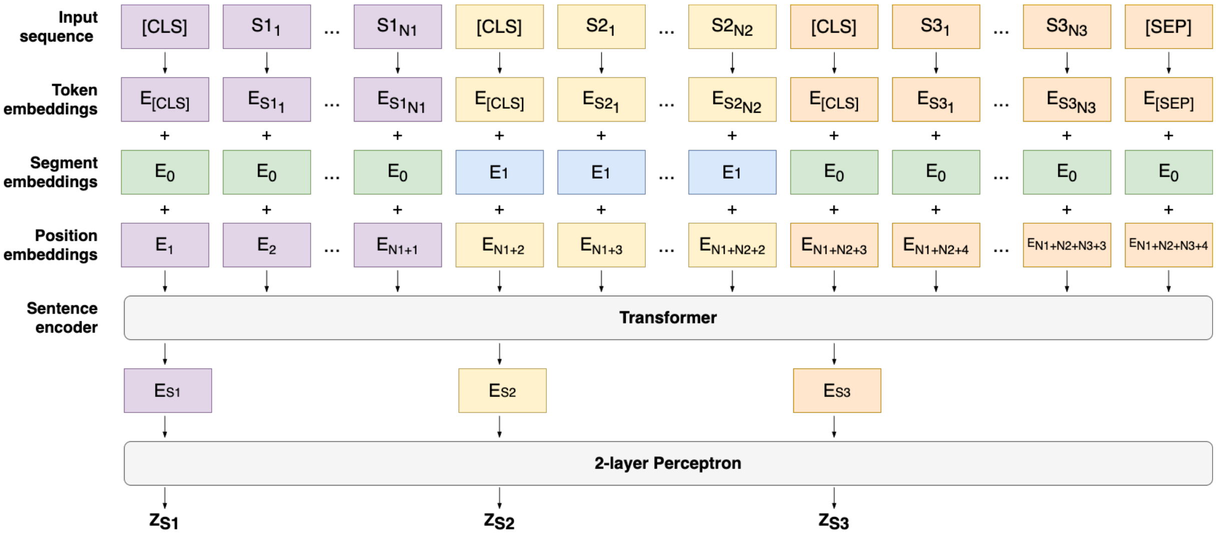

Figure 1: Clean and simple Transformer-based architecture for sentence ordering.

4 Experiments

4.1 Pretrained Models

To study the effect of pretraining objective, language and size on the coherence modeling capacity of the models, we conduct experiments with six pretrained Transformer-based models based on BERT and RoBERTa base models.

BERT (Devlin et al., 2019). Bidirectional Transformer trained with MLM and NSP. Monolingual (16 GB of English data from Book Corpus and Wikipedia). 30k WordPiece vocabulary. 110M parameters.

RoBERTa (Liu et al., 2019). Bidirectional Transformer trained with MLM. Monolingual (160 GB of English data). 50k BPE vocabulary. 125M parameters. Compared with BERT, RoBERTa is trained with dynamic masking (instead of static) on much more data and without NSP loss.

DistilBERT (Sanh et al., 2019). BERT distillation. Monolingual (16 GB of English data from Book Corpus and Wikipedia). 30k WordPiece vocabulary. 66M parameters. 40% less parameters than BERT, while retaining 97% of its language understanding capabilities and being 60% faster.

XLM-R (Conneau et al., 2020). Bidirectional Transformer trained with MLM. Multilingual (2 TB filtered CommonCrawl data; 100 languages). 250k Sentence Piece vocabulary. 270M parameters.

mBERT (Devlin et al., 2019). Bidirectional Transformer trained with MLM and NSP. Multilingual (top 104 languages with the largest Wikipedia). 110k WordPiece vocabulary. 177M parameters.

	Split			Sentences	
Dataset	Train	Dev	Test	Max	Avg
NIPS	2.4K	0.4K	0.4K	15	6
AAN	8.5K	962	2.6K	20	5
NSF	96K	10K	21.5K	40	8.9
ROCStories	78.5K	10K	10k	5	5
SIND	40K	5K	5K	5	5

Table 1: Datasets statistics. Two different domains: scientific papers abstracts (NIPS, AAN, NSF) and storytelling (ROCStories, SIND).

mDistilBERT (Sanh et al., 2019). mBERT distillation. Multilingual (top 104 languages with the largest Wikipedia). 110k WordPiece vocabulary. 134M parameters.

4.2 Datasets for Sentence Ordering

We run our experiments on five datasets from two different domains: scientific papers abstracts (NIPS/AAN/NSF) and storytelling (ROCStories and SIND). We describe all datasets below, and report statistics in Table 1.

Scientific abstracts: NIPS/AAN/NSF abstracts (Logeswaran et al., 2018)). Abstracts from NIPS papers, the ACL Anthology Network corpus and NSF research award papers. The percentage of documents with more than 10 sentences, which will be used to analyse the coherence modeling capabilities of the models on long documents (§5.3), are 1.48%, 2.78% and 24.22%, respectively.

ROCStories (Mostafazadeh et al., 2016). Five-sentence commonsense stories capturing a rich set of causal and temporal commonsense relations between daily event.

SIND (Huang et al., 2016). Sequential vision-to-language dataset containing photo sequences aligned to both descriptive and story language.

Scientific abstracts are challenging due to their specific domain and the higher number of sentences per document (specially high in NSF). ROCStories appears to be the easiest dataset due to the simplicity of its 5-sentence stories. The multimodal nature of SIND makes the dataset very challenging when working only with the textual part, because in the absence of images, the sentence order of many examples is highly ambiguous.

4.3 Metrics

Following previous studies (Prabhumoye et al., 2020; Kumar et al., 2020; Chowdhury et al., 2021; Cui et al., 2020; Zhu et al., 2021a), we use three different metrics:

Sentence accuracy (Acc). Ratio of sentences whose absolute positions are correctly predicted. The metric ranges from 0 (worst) to 100 (best).

Perfect Match Ratio (PMR). Percentage of documents for which the entire order of the sequence is correctly predicted. The metric ranges from 0 (worst) to 100 (best).

Kendall's Tau (τ). Measures how well a ranking agrees with the ground-truth. For a paragraph containing N sentences:

$$\tau = 1 - \frac{2i}{\binom{N}{2}} \tag{3}$$

where i denotes the number of pairs in the predicted sequence with the incorrect relative order (Lapata, 2003). The metric ranges from -1 (worst) to 1 (best).

4.4 Sentence Ordering Baselines

Even though our focus is not on improving the sentence ordering state of the art but on comparing the coherence modeling capabilities of different pretrained Transformer-based models, we compare our results with 5 previous methods for the sake of completeness: Re-BART (Chowdhury et al., 2021), BERSON (Cui et al., 2020), BERT4SO (Zhu et al., 2021a), B-TSort (Prabhumoye et al., 2020) and RankTxNet (Kumar et al., 2020).

4.5 Experimental setup

We use Apache MXNet (Chen et al., 2015) for our experiments, and we train on NVIDIA®V100 Tensor Core GPUs. For the sentence encoder, we rely on the base cased versions of the pretrained models. The Perceptron has two layers with 768 hidden size. We use Adam (Kingma and Ba, 2014) as optimizer, a batch size of 4 and an initial learning rate of $2e^{-6}$, reduced with a polynomial scheduler with 20% of warmup steps. We train to convergence, with 25 patience epochs and a minimum validation accuracy improvement of $1e^{-3}$ afterwards for early-stopping. To mitigate the variance in performance induced by weight initialization and training data order (Dodge et al., 2020), we train each model 3 times with different random seeds and average the results.

5 Results analysis

Table 2 shows a summary of the results on the sentence ordering task for all the models and all the datasets. We will analyse the results along two different axes, comparing 1) monolingual models with their multilingual counterparts, and 2) the three families of models: RoBERTa-based (RoBERTa and XLM-R), BERT-based (BERT and mBERT) and DistilBERT-based (DistilBERT and mDistilBERT). Unless stated otherwise, we will compare the accuracy of the models for simplicity, but the observations hold for all metrics.

As expected, given the simplicity and homogeneity of its 5-sentence documents, all models have a higher accuracy on ROCStories than other datasets. On the other side, SIND and NSF prove to be much more challenging, given the high ambiguity (SIND) and the larger number of sentences of the documents (NSF).

Next, multilingual models do not leverage enough knowledge from other languages to improve their coherence modeling capabilities, and monolingual models clearly outperform their multilingual counterparts in all datasets. DistilBERT-based models show the lower difference between monolingual and multilingual models, with an average accuracy difference of 1.4, followed by BERT-based models, with an accuracy difference of 2.34. RoBERTa-based models show the highest drop, with an accuracy difference of 5.69.

Intuitively, BERT-based models trained with a

	NIPS			AAN			NSF			SIND			ROC		
Models	**Acc**	**Tau**	**PMR**	**Acc**	**Tau**	**PMR**	**Acc**	**Tau**	**PMR**	**Acc**	**Tau**	**PMR**	**Acc**	**Tau**	**PMR**
RoBERTa	62.62	0.82	31.84	72.56	0.86	54.73	37.12	0.67	17.48	55.98	0.64	23.4	82.57	0.89	64.88
BERT	57.32	0.78	27.45	66.69	0.82	46.86	34.35	0.64	15.22	51.63	0.59	18.11	73.89	0.82	48.24
DistilBERT	56.06	0.76	25.12	62.49	0.78	40.99	29.6	0.59	11.32	50.63	0.58	16.69	68.53	0.78	38.95
XLM-R	57.1	0.78	26.95	67.63	0.83	48.05	32.23	0.62	14.1	51.55	0.59	18.04	73.89	0.82	48.65
mBERT	50.54	0.72	19.9	67.07	0.82	47.35	34.21	0.63	15.72	50.33	0.56	16.82	70.02	0.78	42.98
mDistilBERT	53.96	0.75	23.13	62.17	0.78	41.08	30.3	0.57	12.67	48.57	0.55	14.84	65.25	0.75	33.47

Table 2: Sentence ordering results summary. Monolingual models on top, and multilingual models on bottom.

Models	**NIPS**	**AAN**	**NSF**	**SIND**	**ROC**
RoBERTa	85.86	87.83	72.69	74.46	91.77
BERT	83.04	85.33	70.55	70.7	87.5
DistilBERT	82.84	82.91	68.61	70.52	85.16
XLM-R	82.71	85.51	68.78	71.32	87.77
mBERT	77.74	85.59	69.36	69.53	86.15
mDistilBERT	81.55	82.47	66.17	68.3	83.42

Table 3: First and last sentences prediction accuracy.

NSP objective should have an advantage in modeling this task, as predicting if a sentence is the next sentence is a simplification of the sentence ordering task. However, results shows that RoBERTa-based models offer the highest accuracy in both monolingual (RoBERTa) and multilingual (XLM-R) model sets, indicating that indeed the amount and diversity of data compensates the lack of NSP objective. However, RoBERTa outperforms BERT by 5.39 in accuracy in average, while XLM-R outperforms mBERT by only 2.05. This suggests that the advantage gained from the amount of training data is partially mitigated by the exposure of the model to different languages.

DistilBERT-based models, with only 44% of parameters compared to BERT, are able to keep most of the knowledge while offering a faster training and inference times. DistilBERT loses an average of 3.3 accuracy with respect to BERT, and mDistilBERT loses an average of 2.38 accuracy with respect to mBERT. Thus, distilled models are a good choice for applications where efficiency need to be prioritised. Surprisingly, mDistilBERT outperforms mBERT in the NIPS dataset.

5.1 Predicting first and last sentences

The first and last sentences play crucial roles in a paragraph (Oh et al., 2019; Yin et al., 2019), and thus following previous works (Chowdhury et al., 2021; Cui et al., 2020) we report the accuracy of all the models in ordering the first and the last sentences of each document (Table 3).

The accuracy predicting first and last sentences is higher than the overall accuracy shown in Table 2. The difference is wider for NIPS/AAN/NSF, probably due to the presence of strong cues typically marking the last sentence in scientific articles (*Finally*, *To conclude*, ...).

As in the overall results, RoBERTa-based models outperform BERT-based models with a difference of 3.09 accuracy between RoBERTa and BERT and a difference of 1.5 accuracy between XLM-R and mBERT. Again, monolingual models outperform their multilingual counterparts, with the higher difference in RoBERTa-based models (3.3 accuracy), followed by BERT-based models (1.75 accuracy) and DistilBERT-based models (1.63 accuracy). The distilled versions of BERT and mBERT offer a slightly lower accuracy than their original models, with a difference of 1.42 and 1.29 respectively, and mDistilBERT surprisingly outperforms mBERT in the NIPS dataset. Thus, as previously observed, the bigger quantity and diversity of training data seems to be more helpful than sentence-based objectives.

5.2 Sentence displacement analysis

In the sentence ordering task, misplacing a sentence by one position from its original position may not be as harmful to the general coherence as misplacing a sentence by farther positions. To analyse to which degree the models misplace the sentences when ordering documents, we complement our study with an analysis of the displacement of the sentences (Table 4) by calculating the percentage of sentences whose predicted location is within one (*win1*), two (*win2*) or three (*win3*) positions from their original location, following previous studies (Prabhumoye et al., 2020; Cui et al., 2020).

Naturally, *win3* accuracy is closer to 100.00 for those datasets with only five sentences per document (SIND and ROCStories), and lower for those with more sentences per document, with NSF showing the lower values as it is the one containing the

	NIPS			AAN			NSF			SIND			ROC		
Models	**win1**	**win2**	**win3**	**win1**	**win2**	**win3**	**win1**	**win2**	**win3**	**win1**	**win2**	**win3**	**win1**	**win2**	**win3**
RoBERTa	88.27	95.97	98.38	92.03	97.14	98.72	58.99	70.86	78.41	84.89	95.79	99.2	96.57	99.24	99.84
BERT	85.1	93.99	97.32	89.82	96.26	98.51	56.96	69.58	77.68	82.21	94.79	99.06	94.35	98.77	99.81
DistilBERT	85.13	93.63	97.33	87.7	95.64	98.21	54.74	67.77	76.18	81.69	94.68	99.02	92.76	98.41	99.79
XLM-R	86.05	94.64	97.56	90.15	96.52	98.54	54.12	66.59	74.75	82.41	94.93	99.14	94.27	98.75	99.8
mBERT	80.97	92.05	96.54	89.52	96.19	98.48	56.12	68.47	76.43	80.99	94.16	98.83	93.09	98.36	99.78
mDistilBERT	83.24	93.0	97.11	87.41	95.42	98.23	52.04	64.93	73.46	79.96	93.85	98.78	90.84	97.98	99.72

Table 4: Analysis of the displacement of sentences. Percentage of sentences whose predicted location is within one (win1), two (win2) or three (win3) positions from their original location.

	NIPS			AAN			NSF		
Models	**Acc**	**Tau**	**PMR**	**Acc**	**Tau**	**PMR**	**Acc**	**Tau**	**PMR**
RoBERTa	34.8	0.70	0.0	31.17	0.59	0.0	25.03	0.54	1.0
BERT	26.0	0.58	0.0	30.65	0.56	0.0	22.81	0.53	0.63
DistilBERT	25.16	0.63	0.0	27.14	0.54	0.0	20.97	0.49	0.42
XLM-R	31.24	0.63	0.0	31.68	0.58	0.0	19.98	0.47	0.18
mBERT	23.27	0.58	0.0	28.95	0.56	0.0	22.29	0.50	0.69
mDistilBERT	25.58	0.62	0.0	26.52	0.53	0.0	19.1	0.43	0.38

Table 5: Ordering longer documents. Sentence ordering results on documents with 10 sentences or more.

longest documents, with an average of almost 9 sentences per document. On the SIND dataset, there is an average of 30.6 points difference between the general accuracy of the models and the *win1* accuracy. This reinforces our intuition that the SIND dataset is highly ambiguous in order, and the models are prone to swap positions of contiguous sentences.

As previously observed, RoBERTa-based models outperform BERT-base models, monolingual models outperform multilingual models, and full-size models outperform their distillation models. As the window size decreases, the differences get larger, indicating that RoBERTa/monolingual/full-size models offer a lower sentence displacement than their counterparts.

5.3 Performance on longer documents

If the number of sentences in a document is large, these documents will be harder to order. To analyze the models' capabilities of ordering long documents, we evaluate their performance on documents containing more than 10 sentences (Table 5), following previous studies (Prabhumoye et al., 2020; Cui et al., 2020). Indeed, longer documents prove to be much harder to order, and none of the models are able to correctly order any of the long documents, with PMR values dropping to 0 for NIPS and AAN, and less than 1% for NSF.

RoBERTa-based models outperform BERT-base models with an average accuracy difference of 3.85 between RoBERTa and BERT and a difference of 2.8 between XLM-R and mBERT. Compared to the general results in Table 2, RoBERTa-based models lose 47.18% of accuracy when ordering long documents, while BERT-based models lose 50%. Again, monolingual models outperform their multilingual counterparts, but the accuracy loss with respect the general results is very similar for both groups, with and average accuracy loss of 49.18% for monolingual models, and of 49.83% for multilingual models.

As previously observed, the average accuracy difference between the normal and distilled BERT models is higher (2.06) than the difference between the normal and distilled versions of mBERT (1.1), with all models offering similar accuracy losses of around 50% with respect to the general results.

Therefore, in applications where the expected number of sentences per document is large, RoBERTa-based models are a better choice, offering more robust results across sentences lengths.

5.4 Comparison with baselines

Table 6 shows the evaluation metrics of the state-of-the-art methods for sentence ordering (top) along with the evaluation metrics for the models presented in this work (bottom). Despite the simplicity of our method, the model using RoBERTa as encoder ranks in third position for all datasets, and all of our models closely match (and even outperform in some cases) RankTxNet.

	NIPS			AAN			NSF			SIND			ROC		
Models	**Acc**	**Tau**	**PMR**	**Acc**	**Tau**	**PMR**	**Acc**	**Tau**	**PMR**	**Acc**	**Tau**	**PMR**	**Acc**	**Tau**	**PMR**
Re-BART	**77.41**	**0.89**	**57.03**	**84.28**	**0.91**	**73.50**	**50.23**	**0.76**	**29.74**	**64.99**	**0.72**	**43.15**	**90.78**	**0.94**	**81.88**
BERSON	73.87	0.85	48.01	78.03	0.85	59.79	50.02	0.67	23.07	58.91	0.65	31.69	82.86	0.88	68.23
B-TSort	61.48	0.81	32.59	69.22	0.83	50.76	35.21	0.66	10.44	52.23	0.60	20.32	—	—	—
RankTxNet	—	0.75	24.13	—	0.77	39.18	—	0.58	9.78	—	0.57	15.48	—	0.76	38.02
BERT4SO	—	0.78	30.70	—	0.81	45.41	—	0.64	13.00	—	0.59	19.07	—	0.85	55.65
RoBERTa	62.62	0.82	31.84	72.56	0.86	54.73	37.12	0.67	17.48	55.98	0.64	23.4	82.57	0.89	64.88
BERT	57.32	0.78	27.45	66.69	0.82	46.86	34.35	0.64	15.22	51.63	0.59	18.11	73.89	0.82	48.24
DistilBERT	56.06	0.76	25.12	62.49	0.78	40.99	29.6	0.59	11.32	50.63	0.58	16.69	68.53	0.78	38.95
XLM-R	57.1	0.78	26.95	67.63	0.83	48.05	32.23	0.62	14.1	51.55	0.59	18.04	73.89	0.82	48.65
mBERT	50.54	0.72	19.9	67.07	0.82	47.35	34.21	0.63	15.72	50.33	0.56	16.82	70.02	0.78	42.98
mDistilBERT	53.96	0.75	23.13	62.17	0.78	41.08	30.3	0.57	12.67	48.57	0.55	14.84	65.25	0.75	33.47

Table 6: Sentence ordering results summary. The best results are in bold. The last six rows correspond to the models developed in this work, with monolingual models on top, and multilingual models on bottom.

6 Conclusion and Future Work

Our experiments shed light on the coherence modeling capabilities of the monolingual and multilingual versions of BERT, RoBERTa and DistilBERT. We present a simple yet competitive architecture for sentence ordering that relies on pretrained Transformer-based language models as the encoder, and a novel data augmentation strategy designed to use as much knowledge as possible from the data while keeping the training tractable.

We run experiments on 5 different datasets from two different domains: scientific abstracts and commonsense stories. We show that RoBERTa-based models outperform BERT-based models in both monolingual and multilingual subsets, concluding that the intuitive advantage offered by the NSP objective is successfully compensated by the higher amount and diversity of data used to train RoBERTa models. However, the accuracy difference between families is wider in the monolingual set (5.39 accuracy between RoBERTa and BERT, compared to 2.05 accuracy between XLM-R and mBERT). This suggests that the exposure of the models to different languages partially mitigates the advantage gained from the larger and more diverse training data. When trained on multilingual data, RoBERTa-based models lose more accuracy than BERT-based models, being DistilBERT-based models the family with the lower loss. Despite having only 44% of the original parameters, DistilBERT-based models offer a surprisingly strong performance compared with the original models, losing and average accuracy of only 3.31 in the monolingual case, and 2.38 in the multilingual case.

All models offer a higher accuracy when ordering only the first and last sentences of the documents. The difference is especially noticeable for the scientific abstracts datasets (NIPS/AAN/NSF), which are more likely to contain stronger cues such as "*Finally*" or "*To conclude*". Finally, both monolingual and multilingual sets of models are equally affected by the length of the documents, but RoBERTa-based models are more robust, losing around 3% less accuracy when ordering documents with more than 10 sentences.

A question that still remains to be answered is whether the coherence modeling capabilities of RoBERTa-based models can be further improved by the use of different pretraining objectives such as the NSP used in BERT or, on the contrary, the use of larger amounts of more diverse training data and bigger models suffices to leverage the knowledge from Masked Language Modeling alone.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Regina Barzilay and Noemie Elhadad. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 141–148, USA. Association for Computational Linguistics.

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

Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems.

Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi. 2021. Is everything in order? a simple way to order sentences.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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, Brussels, Belgium. Association for Computational Linguistics.

Baiyun Cui, Yingming Li, and Zhongfei Zhang. 2020. BERT-enhanced relational sentence ordering network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6310–6320, Online. 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.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Micha Elsner and Eugene Charniak. 2008. Coreference-inspired coherence modeling. In *Proceedings of ACL-08: HLT, Short Papers*, pages 41–44, Columbus, Ohio. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2011a. 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, Portland, Oregon, USA. Association for Computational Linguistics.

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

Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of COLING 2012*, pages 911–926, Mumbai, India. The COLING 2012 Organizing Committee.

Jingjing Gong, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. End-to-end neural sentence ordering using pointer network. *arXiv preprint arXiv:1611.04953*.

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.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. *arXiv preprint arXiv:1805.06087*.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ioannis Konstas and Mirella Lapata. 2013. Inducing document plans for concept-to-text generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1503–1514.

Pawan Kumar, Dhanajit Brahma, Harish Karnick, and Piyush Rai. 2020. Deep attentive ranking networks for learning to order sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8115–8122.

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.

Mirella Lapata, Regina Barzilay, et al. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, volume 5, pages 1085–1090. Citeseer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for machine reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1694–1704, Melbourne, Australia. Association for Computational Linguistics.

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.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Byungkook Oh, Seungmin Seo, Cheolheon Shin, Eunju Jo, and Kyong-Ho Lee. 2019. Topic-guided coherence modeling for sentence ordering by preserving global and local information. 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 2273–2283, Hong Kong, China. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Tianming Wang and Xiaojun Wan. 2019. Hierarchical attention networks for sentence ordering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7184–7191.

Yuan Wang and Minghe Guo. 2014. A short analysis of discourse coherence. *Journal of Language Teaching and Research*, 5(2):460.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199.

Yongjing Yin, Fandong Meng, Jinsong Su, Yubin Ge, Lingeng Song, Jie Zhou, and Jiebo Luo. 2020. Enhancing pointer network for sentence ordering with pairwise ordering predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9482–9489.

Yongjing Yin, Linfeng Song, Jinsong Su, Jiali Zeng, Chulun Zhou, and Jiebo Luo. 2019. Graph-based neural sentence ordering. *arXiv preprint arXiv:1912.07225*, pages 5387–5393.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations*.

Yutao Zhu, Jian-Yun Nie, Kun Zhou, Shengchao Liu, Yabo Ling, and Pan Du. 2021a. Bert4so: Neural sentence ordering by fine-tuning bert. *arXiv preprint arXiv:2103.13584*.

Yutao Zhu, Kun Zhou, Jian-Yun Nie, Shengchao Liu, and Zhicheng Dou. 2021b. Neural sentence ordering based on constraint graphs.