Discourse Probing of Pretrained Language Models

Fajri Koto **Jey Han Lau** **Timothy Baldwin**
School of Computing and Information Systems
The University of Melbourne
ffajri@student.unimelb.edu.au, jeyhan.lau@gmail.com, tbaldwin@unimelb.edu.au

Abstract

Existing work on probing of pretrained language models (LMs) has predominantly focused on sentence-level syntactic tasks. In this paper, we introduce document-level discourse probing to evaluate the ability of pretrained LMs to capture document-level relations. We experiment with 7 pretrained LMs, 4 languages, and 7 discourse probing tasks, and find BART to be overall the best model at capturing discourse — but only in its encoder, with BERT performing surprisingly well as the baseline model. Across the different models, there are substantial differences in which layers best capture discourse information, and large disparities between models.

1 Introduction

The remarkable development of pretrained language models (Devlin et al., 2019; Lewis et al., 2020; Lan et al., 2020) has raised questions about what precise aspects of language these models do and do not capture. Probing tasks offer a means to perform fine-grained analysis of the capabilities of such models, but most existing work has focused on sentence-level analysis such as syntax (Hewitt and Manning, 2019; Jawahar et al., 2019; de Vries et al., 2020), entities/relations (Papanikolaou et al., 2019), and ontological knowledge (Michael et al., 2020). Less is known about how well such models capture broader discourse in documents.

Rhetorical Structure Theory is a framework for capturing how sentences are connected and describing the overall structure of a document (Mann and Thompson, 1986). A number of studies have used pretrained models to classify discourse markers (Sileo et al., 2019) and discourse relations (Nie et al., 2019; Shi and Demberg, 2019), but few (Koto et al., to appear) have systematically investigated the ability of pretrained models to model discourse structure. Furthermore, existing work relating to discourse probing has typically focused exclusively on the BERT-base model, leaving open the question of how well these findings generalize to other models with different pretraining objectives, for different languages, and different model sizes.

Model	Type	#Param	#Data	Objective
BERT	Enc	110M	16GB	MLM+NSP
RoBERTa		110M	160GB	MLM
ALBERT		12M	16GB	MLM+SOP
ELECTRA		110M	16GB	MLM+DISC
GPT-2	Dec	117M	40GB	LM
BART	Enc+Dec	121M	160GB	DAE
T5		110M	750GB	DAE

Table 1: Summary of all English pretrained language models used in this work. "MLM" = masked language model, "NSP" = next sentence prediction, "SOP" = sentence order prediction, "LM" = language model, "DISC" = discriminator, and "DAE" = denoising autoencoder.

Our research question in this paper is: *How much discourse structure do layers of different pretrained language models capture, and do the findings generalize across languages?*

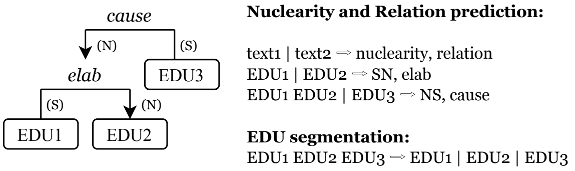
There are two contemporaneous related studies that have examined discourse modelling in pretrained language models. Upadhye et al. (2020) analyzed how well two pretrained models capture referential biases of different classes of English verbs. Zhu et al. (2020) applied the model of Feng and Hirst (2014) to parse IMDB documents (Maas et al., 2011) into discourse trees. Using this (potentially noisy) data, probing tasks were conducted by mapping attention layers into single vectors of document-level rhetorical features. These features, however, are unlikely to capture all the intricacies of inter-sentential abstraction as their input is formed based on discourse relations[1] and aggregate statistics on the distribution of discourse units.

[1] For example, they only consider discourse relation labels and ignore nuclearity.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3849–3864
June 6–11, 2021. ©2021 Association for Computational Linguistics

Probing Task	English	Chinese	German	Spanish
(1) 4-way NSP (2) Sentence Ordering	XSUM articles (Narayan et al., 2018) Split: 8K/1K/1K	Wikipedia (ZH) Split: 8K/1K/1K	Wikipedia (DE) Split: 8K/1K/1K	Wikipedia (ES) Split: 8K/1K/1K
(3) Discourse Connective	Sampled DisSent dataset (Nie et al., 2019) #Labels: 15 Split: 10K/1K/1K	CDTB (Li et al., 2014) #Labels: 22 Split: 1539/76/168	Potsdam Commentary (Bourgonje and Stede, 2020) #Labels: 15 Split: 900/148/159	N/A
(4) RST Nuclearity (5) RST Relation	RST-DT (Carlson et al., 2001) #Labels (nuc/rel): 3/18 Split: 16903/1943/2308	CDTB (Li et al., 2014) #Labels (nuc/rel): 3/4 Split: 6159/353/809	Potsdam Commentary (Bourgonje and Stede, 2020) #Labels (nuc/rel): 3/31 Split: 1892/289/355	RST-Spanish Treebank (da Cunha et al., 2011) #Labels (nuc/rel): 3/29 Split: 2042/307/421
(6) RST EDU Segmentation	RST-DT (Carlson et al., 2001) Split: 312/35/38 docs	CDTB (Li et al., 2014) Split: 2135/105/241 p'graphs	Potsdam Commentary (Bourgonje and Stede, 2020) Split: 131/20/25 docs	RST-Spanish Treebank (da Cunha et al., 2011) Split: 200/34/30 docs
(7) Cloze Story Test	(Mostafazadeh et al., 2016) Split: 1683/188/1871	N/A	N/A	N/A

Table 2: A summary of probing tasks and datasets for each of the four languages. "Split" indicates the number of train/development/test instances.



Nuclearity and Relation prediction:

text1 | text2 → nuclearity, relation
EDU1 | EDU2 → SN, elab
EDU1 EDU2 | EDU3 → NS, cause

EDU segmentation:
EDU1 EDU2 EDU3 → EDU1 | EDU2 | EDU3

Figure 1: Illustration of the RST discourse probing tasks (Tasks 4–6).

To summarize, we introduce 7 discourse-related probing tasks, which we use to analyze 7 pretrained language models over 4 languages: English, Mandarin Chinese, German, and Spanish. Code and public-domain data associated with this research is available at https://github.com/fajri91/discourse_probing.

2 Pretrained Language Models

We outline the 7 pretrained models in Table 1. They comprise 4 encoder-only models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and ELECTRA (Clark et al., 2020); 1 decoder-only model: GPT-2 (Radford et al., 2019); and 2 encoder–decoder models: BART (Lewis et al., 2020) and T5 (Raffel et al., 2019). To reduce the confound of model size, we use pretrained models of similar size (∼110m model parameters), with the exception of ALBERT which is designed to be lighter weight. All models have 12 transformer layers in total; for BART and T5, this means their encoder and decoder have 6 layers each. Further details of the models are provided in the Supplementary Material.

3 Probing Tasks for Discourse Coherence

We experiment with a total of seven probing tasks, as detailed below. Tasks 4–6 are component tasks of discourse parsing based on rhetorical structure theory (RST; Mann and Thompson (1986)). In an RST discourse tree, EDUs are typically clauses or sentences, and are hierarchically connected with discourse labels denoting: (1) nuclearity = nucleus (N) vs. satellite (S);[2] and (2) discourse relations (e.g. *elaborate*). An example of a binarized RST discourse tree is given in Figure 1.

1. **Next sentence prediction.** Similar to the next sentence prediction (NSP) objective in BERT pretraining, but here we frame it as a 4-way classification task, with one positive and 3 negative candidates for the next sentence. The preceding context takes the form of between 2 and 8 sentences, but the candidates are always single sentences.

2. **Sentence ordering.** We shuffle 3–7 sentences and attempt to reproduce the original order. This task is based on Barzilay and Lapata (2008) and Koto et al. (2020), and is assessed based on rank correlation relative to the original order.

3. **Discourse connective prediction.** Given two sentences/clauses, the task is to identify an appropriate discourse marker, such as *while*,

[2] The satellite is a supporting EDU for the nucleus.

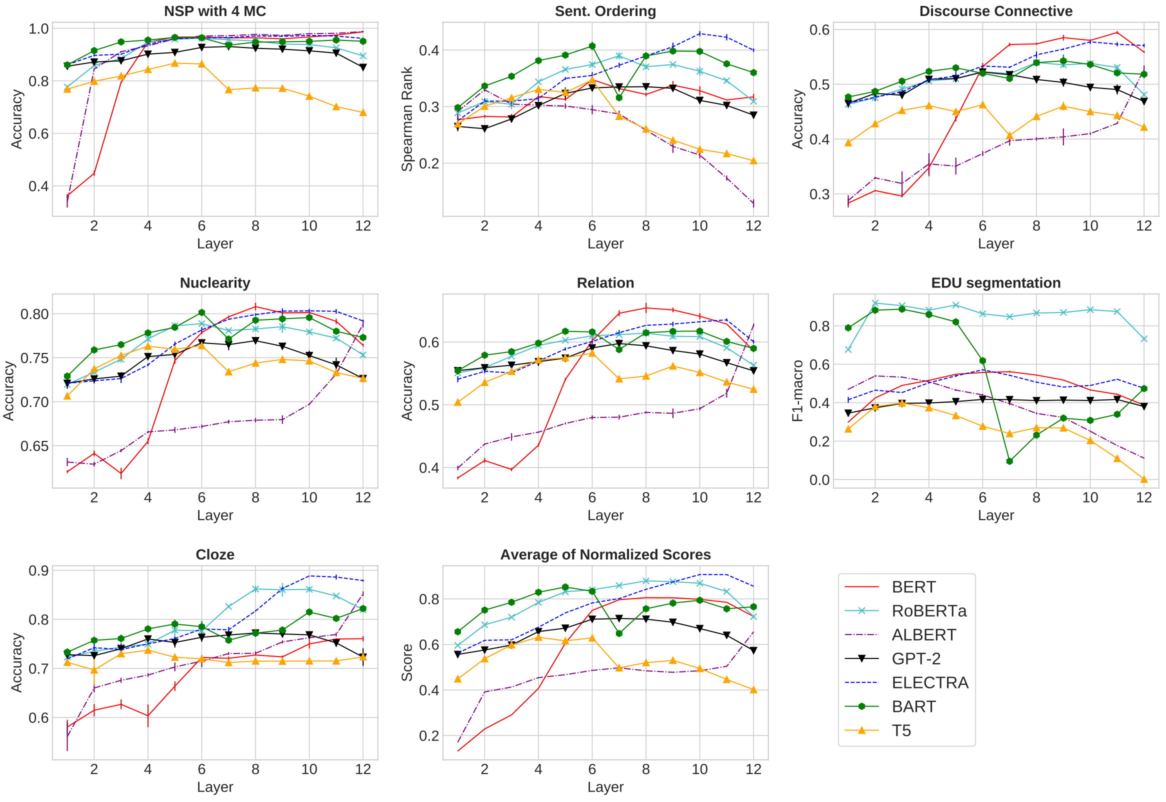


Figure 2: Probing task performance on English for each of the seven tasks, plus the average across all tasks. For BART and T5, layers 7–12 are the decoder layers. All results are averaged over three runs, and the vertical line for each data point denotes the standard deviation (noting that most results have low s.d., meaning the bar is often not visible).

or, or *although* (Nie et al., 2019), representing the conceptual relation between the sentences/clauses.

4. **RST nuclearity prediction.** For a given ordered pairing of (potentially complex) EDUs which are connected by an unspecified relation, predict the nucleus/satellite status of each (see Figure 1).

5. **RST relation prediction.** For a given ordered pairing of (potentially complex) EDUs which are connected by an unspecified relation, predict the relation that holds between them (see Figure 1).

6. **RST elementary discourse unit (EDU) segmentation.** Chunk a concatenated sequence of EDUs into its component EDUs.

7. **Cloze story test.** Given a 4-sentence story context, pick the best ending from two possible options (Mostafazadeh et al., 2016; Sharma et al., 2018). This task is harder than NSP, as it requires an understanding of commonsense and storytelling (Chaturvedi et al., 2017; Liu et al., 2018).

4 Experimental Setup

We summarize all data (sources, number of labels, and data split) in Table 2. This includes English, Chinese, German, and Spanish for each probing task. For NSP and sentence ordering, we generate data from news articles and Wikipedia. For the RST tasks, we use discourse treebanks for each of the four languages.

We formulate all probing tasks except sentence ordering and EDU segmentation as a classification problem, and evaluate using accuracy. During fine-tuning, we add an MLP layer on top of the pretrained model for classification, and only update the MLP parameters (all other layers are frozen). We use the [CLS] embedding for BERT and ALBERT following standard practice, while for other models we perform average pooling to obtain a vector for each sentence, and concatenate them as the input to the MLP.[3]

[3] BERT and ALBERT performance with average pooling

For sentence ordering, we follow Koto et al. (2020) and frame it as a sentence-level sequence labelling task, where the goal is to estimate $P(r|s)$, where r is the rank position and s the sentence. The task has 7 classes, as we have 3–7 sentences (see Section 3). At test time, we choose the label sequence that maximizes the sequence probability. Sentence embeddings are obtained by average pooling. The EDU segmentation task is also framed as a binary sequence labelling task (segment boundary or not) at the (sub)word level. We use Spearman rank correlation and macro-averaged F1 score to evaluate sentence ordering and EDU segmentation, respectively.

We use a learning rate $1e-3$, warm-up of 10% of total steps, and the development set for early stopping in all experiments. All presented results are averaged over three runs.[4]

5 Results and Analysis

In Figure 2, we present the probing task performance on English for all models based on a representation generated from each of the 12 layers of the model. First, we observe that most performance fluctuates (non-monotonic) across layers except for some models in the NSP task and some ALBERT results in the other probing tasks. We also found that most models except ALBERT tend to have a very low standard deviation based on three runs with different random seeds.

We discover that all models except T5 and early layers of BERT and ALBERT perform well over the NSP task, with accuracy ≥ 0.8, implying it is a simple task. However, they all struggle at sentence ordering (topping out at $\rho \sim 0.4$), suggesting that they are ineffective at modelling discourse over multiple sentences; this is borne out in Figure 4, where performance degrades as the number of sentences to re-order increases.

Interestingly, for Discourse Connectives, RST Nuclearity, and RST Relation Prediction, the models produce similar patterns, even though the discourse connective data is derived from a different dataset and theoretically divorced from RST. BART outperforms most other models in layers 1–6 for these tasks (a similar observation is found for NSP and Sentence Ordering) with BERT and ALBERT struggling particularly in the earlier layers. For EDU segmentation, RoBERTa and again the first few layers of BART perform best. For the Cloze Story Test, all models seem to improve as we go deeper into the layers, suggesting that high-level story understanding is captured deeper in the models.

We summarize the overall performance by calculating the averaged normalized scores in the last plot in Figure 2.[5] RoBERTa and BART appear to be the best overall models at capturing discourse information, but only in the *encoder layers* (the first 6 layers) for BART. We hypothesize that the BART decoder focuses on sequence generation, and as such is less adept at language understanding. This is supported by a similar trend for T5, also a denoising autoencoder. BERT does surprisingly well (given that it's the baseline model), but mostly in the deeper layers (7–10), while ELECTRA performs best at the three last layers.

In terms of the influence of training data, we see mixed results. BART and RoBERTa are the two best models, and both are trained with more data than most models (an order of magnitude more; see Table 1). But T5 (and to a certain extent GPT-2) are also trained with more data (in fact T5 has the most training data), but their discourse modelling performance is underwhelming. In terms of training objectives, it appears that a pure decoder with an LM objective (GPT-2) is less effective at capturing discourse structure. ALBERT, the smallest model (an order of magnitude less parameters than most), performs surprisingly well (with high standard deviation), but only at its last layer, suggesting that discourse knowledge is concentrated deep inside the model.

Lastly, we explore whether these trends hold if we use a larger model (BERT-base vs. BERT-large) and for different languages (again based on monolingual BERT models for the respective languages). Results are presented in Figure 3. For model size ("English (large)" vs. "English"), the overall pattern is remarkably similar, with a slight uplift in absolute results with the larger model. Between the 4 different languages (English, Chinese, German, and Spanish), performance varies for all tasks except for NSP (e.g. EDU segmentation appears to be easiest in Chinese, and relation prediction is the hardest in German), but the *shape* of the lines is largely the same, indicating the optimal layers for

is in included in the Appendix.

[4] More details of the training configuration are given in the Appendix.

[5] Given a task, we perform min–max normalization for all model-layer scores (7×12 scores in total), and then compute the average over all tasks for each model's layer.

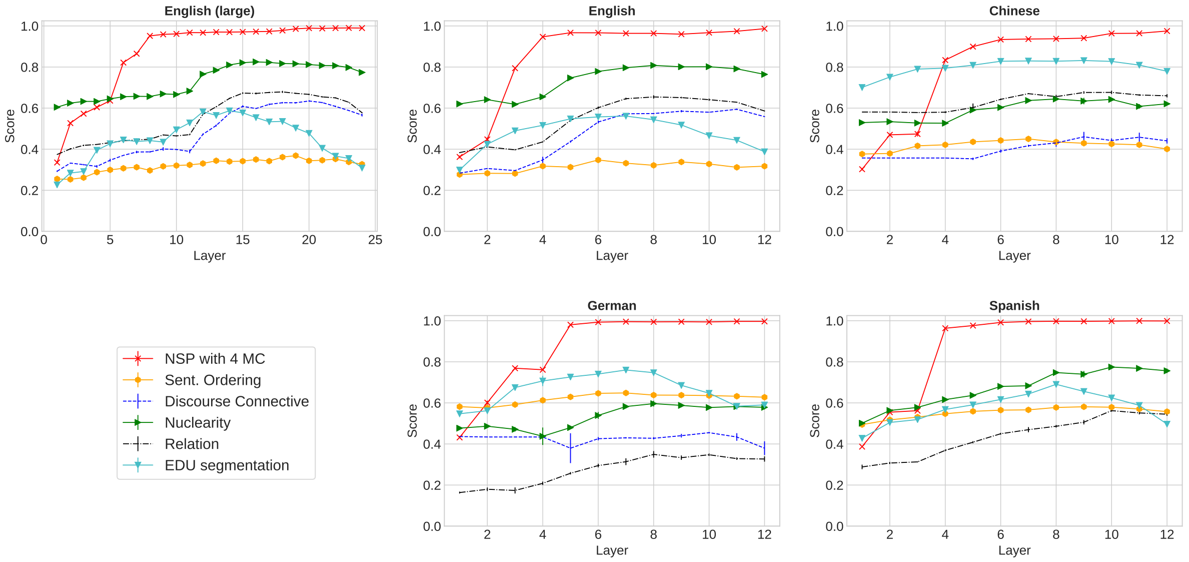


Figure 3: Discourse performance of BERT across different languages. All results are averaged over three runs, and a vertical line is used to denote the standard deviation for each data point (most of which are not visible, due to the low standard deviation).

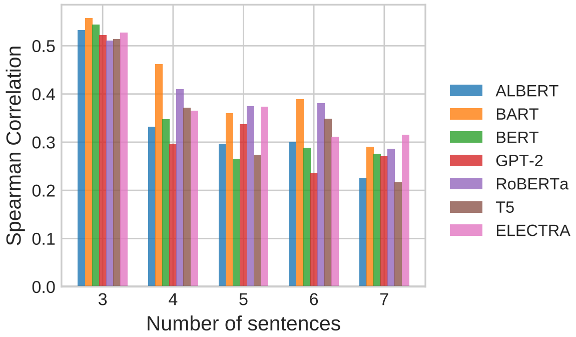


Figure 4: Sentence ordering task breakdown based on the best layer of each model.

a particular task are consistent across languages.

6 Conclusion

We perform probing on 7 pretrained language models across 4 languages to investigate what discourse effects they capture. We find that BART's encoder and RoBERTa perform best, while pure language models (GPT-2) struggle. Interestingly, we see a consistent pattern across different languages and model sizes, suggesting that the trends we found are robust across these dimensions.

Acknowledgements

We are grateful to the anonymous reviewers for their helpful feedback and suggestions. The first author is supported by the Australia Awards Scholarship (AAS), funded by the Department of Foreign Affairs and Trade (DFAT), Australia. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at The University of Melbourne. This facility was established with the assistance of LIEF Grant LE170100200.

References

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

Peter Bourgonje and Manfred Stede. 2020. The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1061–1066, Marseille, France. European Language Resources Association.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1603–1614, Copenhagen, Denmark. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining text encoders as discriminators rather than generators. In *ICLR 2020: Eighth International Conference on Learning Representations*.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST

Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? A closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, 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.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521, Baltimore, Maryland. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. to appear. Top-down discourse parsing via sequence labelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*.

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 *ICLR 2020: Eighth International Conference on Learning Representations*.

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.

Yancui Li, Wenhe Feng, Jing Sun, Fang Kong, and Guodong Zhou. 2014. Building Chinese discourse corpus with connective-driven dependency tree structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2105–2114, Doha, Qatar. Association for Computational Linguistics.

Fei Liu, Trevor Cohn, and Timothy Baldwin. 2018. Narrative modeling with memory chains and semantic supervision. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 278–284, 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. *arXiv preprint arXiv:1907.11692*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1986. Assertions from discourse structure. In *Strategic Computing - Natural Language Workshop: Proceedings of a Workshop Held at Marina del Rey, California, May 1-2, 1986*.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812, Online. Association for Computational Linguistics.

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

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Yannis Papanikolaou, Ian Roberts, and Andrea Pierleoni. 2019. Deep bidirectional transformers for relation extraction without supervision. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 67–75, Hong Kong, China. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rishi Sharma, James Allen, Omid Bakhshandeh, and Nasrin Mostafazadeh. 2018. Tackling the story ending biases in the story cloze test. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 752–757, Melbourne, Australia. Association for Computational Linguistics.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. 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 5790–5796, Hong Kong, China. Association for Computational Linguistics.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. 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 3477–3486, Minneapolis, Minnesota. Association for Computational Linguistics.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 2020. Predicting reference: What do language models learn about discourse models? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 977–982, Online. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. 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 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Zining Zhu, Chuer Pan, Mohamed Abdalla, and Frank Rudzicz. 2020. Examining the rhetorical capacities of neural language models. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 16–32, Online. Association for Computational Linguistics.

A Pretrained Language Models

The pretrained models are sourced from Huggingface (https://huggingface.co/), as detailed in Tables 3 and 4.

Model	Huggingface model
BERT	`bert-base-uncased`
BERT (large)	`bert-large-uncased`
RoBERTa	`roberta-base`
ALBERT	`albert-base-v2`
ELECTRA	`electra-base-discriminator`
GPT-2	`gpt2`
BART	`bart-base`
T5	`t5-small`

Table 3: List of English pretrained language models

Language	Huggingface model
Chinese	`bert-base-chinese`
German	`bert-base-german-dbmdz-uncased`
Spanish	`bert-base-spanish-wwm-uncased`

Table 4: List of non-English BERT models.

B Data Construction, Examples, and Training Configuration

B.1 Next Sentence Prediction

We use spaCy (https://spacy.io/) to perform sentence tokenization, and ensure that the distractor options in the training set do not overlap with the test set. For all languages and models, the training configurations are similar: the maximum tokens in the context and the next sentence are 450 and 50, respectively. If the token lengths are more than this, we truncate the context from the beginning of the sequence, and truncate the next sentence at the end of the sequence. We concatenate context with each option, and perform binary classification.

Other training configuration details: learning rate = 1e-3, Adam epsilon =1e-8, maximum gradient norm = 1.0, maximum epochs = 20, warmup = 10% of the training steps, and patience for early stopping = 5 epochs.

B.2 Sentence Ordering

In generating sentence ordering data, we once again use spaCy (https://spacy.io/) to perform sentence tokenization. For all languages and models, the training configurations are similar, with the maximum tokens in each sentence = 50, learning rate = 1e-3, Adam epsilon = 1e-8, maximum gradient norm = 1.0, training epochs = 20, warmup = 10% of the training steps, and patience for early stopping = 10 epochs.

#Sentence (context)	Total
2	2500
4	2500
6	2500
8	2500
Total	10000

Table 5: NSP data based on the number of sentences.

Context
s1: The Eastern Star, mostly carrying elderly tourists, capsized on 1 June near Jianli in Hubei province. **s2:** Just 14 of the 456 passengers and crew are known to have survived.
Options
0: The channel recently said its signal was carried by 22 satellites **0:** That step has become a huge challenge for opposition candidates **0:** Six men were convicted and then acquitted of the atrocity and no-one has since been convicted of involvement in the bombing **1:** A search is continuing for eight people who remain missing.

Table 6: Example of English NSP data with 2-sentence context. 1 indicates the correct next sentence.

#Sentence	Total
3	2000
4	2000
5	2000
6	2000
7	2000
Total	10000

Table 7: Sentence ordering data based on number of sentence.

B.3 Discourse Connective Prediction

As our Chinese and German data is extracted from discourse treebanks, the number of distinct connective words varies. For instance, in the Chinese discourse treebank, we find 246 unique connective

Context
s0: West Mercia Police said the police do not encourage members of the public to pursue their own investigations. **s1:** David John Poole, from Hereford, poses online as a 14-year-old girl and says he has been sent hundreds of explicit messages. **s2:** He says his work has led to two arrests in four weeks.
Correct order: 2–0–1

Table 8: Example of English sentence ordering data

words. To simplify this, we set the connective word to *OTHER* if its word frequency is less than 12.

For all languages and models, the training configurations are: maximum token length of each sentence = 50, learning rate = 1e-3, Adam epsilon = 1e-8, maximum gradient norm = 1.0, maximum epochs = 20, warmup = 10% of the training steps, and patience for early stopping = 10 epochs.

ENGLISH

Train:
but: 2237, and: 2190, as: 1547, when: 1085, if: 993, before: 462, while: 358, because: 335, though: 229, after: 196, so: 180, although: 84, still: 38, then: 35, also: 31

Development:
but: 222, and: 192, as: 181, when: 116, if: 104, because: 44, before: 40, while: 28, though: 28, so: 16, after: 16, also: 5, although: 5, then: 2, still: 1

Test:
and: 211, but: 202, as: 153, when: 129, if: 97, before: 48, while: 44, because: 41, after: 20, though: 19, although: 11, so: 9, still: 7, also: 5, then: 4

CHINESE

Train:
other: 520, 并: 182, 其中: 131, 也: 118, 但: 60, 而: 60, 还: 55, 以: 47, 使: 43, 后: 42, 为: 41, 同时: 37, 由于: 34, 因此: 28, 如: 26, 又: 20, 为了: 19, 如果: 17, 而且: 16, 但是: 15, 因为: 15, 虽然-但: 13

Development:
other: 22, 并: 7, 也: 6, 而: 6, 其中: 4, 但: 4, 因为: 4, 为: 3, 还: 3, 而且: 3, 又: 2, 如果: 2, 同时: 2, 使: 2, 后: 2, 如: 1, 由于: 1, 虽然-但: 1, 为了: 1

Test:
other: 60, 其中: 18, 并: 18, 也: 10, 使: 10, 还: 9, 同时: 8, 而: 6, 以: 5, 但: 5, 为: 4, 又: 3, 因为: 2, 虽然-但: 2, 由于: 2, 为了: 2, 因此: 2, 而且: 1, 如: 1

GERMAN

Train:
other: 336, und: 191, doch: 62, wenn: 56, aber: 56, denn: 36, dann: 23, auch: 23, sondern: 19, oder: 19, so: 18, also: 17, deshalb: 16, weil: 15, als: 13

Development:
other: 50, und: 32, doch: 12, wenn: 11, aber: 9, denn: 6, dann: 5, so: 5, auch: 5, oder: 4, deshalb: 3, weil: 3, sondern: 2, als: 1

Test:
other: 69, und: 23, doch: 11, aber: 10, denn: 9, wenn: 8, dann: 8, so: 4, weil: 4, auch: 4, sondern: 3, deshalb: 3, oder: 2, als: 1

Figure 5: Discourse connective word distribution.

B.4 RST-related Tasks

In Figures 7 and 8, we present the distribution of the nuclearity and relation labels for the 4 different discourse treebanks. The English treebank is significantly larger, with a strong preference for the NS (nuclear–satellite) relationship. Unlike other languages, the proportion of NN (nuclear–nuclear) relationships in the Chinese discourse treebank (CDTB) is the highest. We also notice that the relation label set in CDTB is the simplest, with only 4 labels.

Most of the training details for nuclearity and relation prediction are the same as for the NSP task, except we set the maximum token length of each sentence to 250. Particularly for EDU segmentation, we set the maximum token length in a document to 512.

B.5 Cloze Story Test

As discussed in Table 2, we use cloze story test version-1 (Mostafazadeh et al., 2016). Although version-2 (Sharma et al., 2018) is better in terms of story biases, the gold labels for the test set are not publicly available, which limited our ability to explore different layers of a broad range of pretrained language models (due to rate limiting of test evaluation).

For the data split, we followed previous work (Liu et al., 2018) in splitting the development set into a training and validation set. We perform binary classification similar to the NSP task, by first merging all 4-sentence stories into a single text (context). We limit the context to a maximum of 450 tokens, and each candidate sentence (as the story ending) is limited to 50 tokens. Other training details are the same as for the NSP task.

ENGLISH
S1: Two men nudged the door open. **S2:** Slipped into the room with him. **Connective word:** and
CHINESE
S1: 目前，约有十五万家外商投资企业在中国银行开立帐户， **S2:** 二万多家获得中国银行的贷款支持。 **Connective word:** 其中
GERMAN
S1: der mann bezahlte viele handwerker nicht **S2:** wurde voriges jahr zu einer mehrjährigen haftstrafe verurteilt **Connective word:** und

Figure 6: Discourse connective: data examples

ENGLISH
Train: NS: 10348, NN: 3853, SN: 2702 **Development:** NS: 1195, NN: 407, SN: 341 **Test:** NS: 1373, NN: 507, SN: 428
CHINESE
Train: NN: 3133, NS: 1784, SN: 1242 **Development:** NN: 432, NS: 219, SN: 158 **Test:** NN: 188, NS: 107, SN: 58
GERMAN
Train: SN: 752, NS: 733, NN: 407 **Development:** NS: 116, SN: 106, NN: 67 **Test:** NS: 150, SN: 133, NN: 72
SPANISH
Train: NS: 1011, SN: 570, NN: 461 **Development:** NS: 163, SN: 73, NN: 71 **Test: NS**: 211, SN: 121, NN: 89

Figure 7: Nuclearity label distribution.

ENGLISH
elab: 7830, attr: 3041, list: 1957, same: 1390, cont: 1108, evid: 967, back: 931, cause: 685, eval: 588, purp: 560, temp: 526, cond: 326, comp: 299, mann: 225, summ: 222, topic: 204, prob: 153, text: 142
CHINESE
并列类: 4144, 解说类: 1630, 因果类: 1333, 转折类: 214
GERMAN
reason: 267, interpretation: 232, elaboration: 204, joint: 203, background: 163, list: 138, concession: 125, antithesis: 123, conjunction: 117, condition: 116, circumstance: 113, e-elaboration: 111, cause: 101, evidence: 99, preparation: 87, evaluation-s: 80, contrast: 49, result: 46, evaluation-n: 38, purpose: 30, sequence: 29, restatement: 17, means: 11, disjunction: 10, summary: 9, solutionhood: 7, justify: 4, otherwise: 3, enablement: 2, unless: 1, motivation: 1
SPANISH
elaboración: 625, preparación: 370, lista: 257, fondo: 178, unión: 168, medio: 135, resultado: 134, circunstancia: 122, propósito: 115, secuencia: 79, interpretación: 77, antítesis: 67, contraste: 61, causa: 57, evidencia: 49, condición: 47, concesión: 44, justificación: 39, same-unit: 33, solución: 26, motivación: 21, reformulación: 16, conjunción: 14, disyunción: 9, evaluación: 9, resumen: 8, capacitación: 5, alternativa: 3, unless: 2

Figure 8: Relation label distribution.

C Full Experimental Results

Layer	NSP	Sent. Ord.	Discourse Conn.	Nuclearity	Relation	EDU segment.	Cloze ST.
BERT (English); std = 0.00 – 0.02							
1	0.36	0.28	0.28	0.62	0.38	0.30	0.58
2	0.45	0.28	0.31	0.64	0.41	0.42	0.61
3	0.79	0.28	0.30	0.62	0.40	0.49	0.63
4	0.95	0.32	0.35	0.65	0.44	0.52	0.60
5	0.97	0.31	0.44	0.75	0.54	0.55	0.66
6	0.97	**0.35**	0.53	0.78	0.60	**0.56**	0.72
7	0.96	0.33	0.57	0.80	**0.65**	**0.56**	0.72
8	0.96	0.32	0.57	**0.81**	**0.65**	0.54	0.73
9	0.96	0.34	**0.59**	0.80	**0.65**	0.52	0.72
10	0.97	0.33	0.58	0.80	0.64	0.47	0.75
11	0.97	0.31	**0.59**	0.79	0.63	0.44	**0.76**
12	**0.99**	0.32	0.56	0.76	0.59	0.39	**0.76**
RoBERTa (English); std = 0.00 – 0.02							
1	0.78	0.29	0.46	0.72	0.55	0.68	0.72
2	0.86	0.31	0.48	0.73	0.56	**0.92**	0.73
3	0.88	0.30	0.49	0.75	0.58	0.90	0.74
4	0.95	0.34	0.51	0.77	0.59	0.88	0.75
5	**0.96**	0.37	0.51	**0.79**	0.60	0.91	0.78
6	**0.96**	0.37	0.52	**0.79**	**0.61**	0.86	0.78
7	**0.96**	**0.39**	0.52	0.78	**0.61**	0.85	0.83
8	0.95	0.37	**0.54**	0.78	**0.61**	0.87	**0.86**
9	0.94	0.37	**0.54**	**0.79**	**0.61**	0.87	**0.86**
10	0.94	0.36	**0.54**	0.78	**0.61**	0.88	**0.86**
11	0.93	0.35	0.53	0.77	0.59	0.87	0.85
12	0.90	0.31	0.48	0.75	0.56	0.73	0.82
ALBERT (English); std = 0.00 – 0.03							
1	0.34	0.29	0.29	0.63	0.40	0.47	0.56
2	0.85	**0.33**	0.33	0.63	0.44	**0.54**	0.66
3	0.91	0.30	0.32	0.64	0.45	0.53	0.68
4	0.93	0.30	0.35	0.67	0.46	0.51	0.69
5	0.96	0.30	0.35	0.67	0.47	0.47	0.70
6	0.97	0.29	0.37	0.67	0.48	0.44	0.71
7	0.97	0.29	0.40	0.68	0.48	0.40	0.73
8	0.98	0.26	0.40	0.68	0.49	0.34	0.73
9	0.97	0.23	0.40	0.68	0.49	0.32	0.75
10	0.98	0.21	0.41	0.70	0.49	0.25	0.76
11	0.98	0.17	0.43	0.73	0.52	0.18	0.77
12	**0.99**	0.13	**0.53**	**0.79**	**0.63**	0.11	**0.85**
ELECTRA (English); std = 0.00 – 0.02							
1	0.86	0.27	0.47	0.72	0.54	0.42	0.72
2	0.90	0.31	0.48	0.72	0.55	0.47	0.74
3	0.90	0.31	0.49	0.73	0.55	0.45	0.74
4	0.94	0.31	0.51	0.74	0.57	0.50	0.75
5	0.96	0.35	0.52	0.77	0.59	0.54	0.76
6	0.96	0.36	0.53	0.78	0.60	**0.57**	0.78
7	0.96	0.37	0.53	0.79	0.61	0.54	0.78
8	**0.97**	0.39	0.55	**0.80**	0.63	0.51	0.82
9	**0.97**	0.41	0.56	**0.80**	0.63	0.48	0.86
10	**0.97**	**0.43**	**0.58**	**0.80**	0.63	0.49	**0.89**
11	**0.97**	0.42	0.57	**0.80**	**0.64**	0.52	**0.89**
12	0.96	0.40	0.57	0.79	0.60	0.48	0.88

Table 9: Full results for BERT, RoBERTa, ALBERT, and ELECTRA over English.

Layer	**NSP**	**Sent.**	**Discourse**	**Nuclearity**	**Relation**	**EDU**	**Cloze**
		Ord.	**Conn.**			**segment.**	**ST.**
GPT-2 (English); std = 0.00 – 0.02							
1	0.86	0.26	0.47	0.72	0.55	0.35	0.73
2	0.87	0.26	0.48	0.73	0.56	0.37	0.73
3	0.88	0.28	0.48	0.73	0.56	0.40	0.74
4	0.90	0.30	0.51	0.75	0.57	0.40	0.76
5	0.91	0.32	0.51	0.75	0.57	0.41	0.75
6	**0.93**	0.33	**0.52**	**0.77**	0.59	**0.42**	0.76
7	**0.93**	0.33	**0.52**	0.76	**0.60**	**0.42**	**0.77**
8	0.92	**0.34**	0.51	**0.77**	0.59	0.41	**0.77**
9	0.92	0.33	0.50	0.76	0.59	0.41	**0.77**
10	0.91	0.31	0.49	0.75	0.58	0.41	**0.77**
11	0.91	0.30	0.49	0.74	0.57	0.42	0.75
12	0.85	0.28	0.47	0.73	0.55	0.38	0.72
BART (English); Layers 7–12 are the decoder; std = 0.00 – 0.01.							
1	0.86	0.30	0.48	0.73	0.55	0.79	0.73
2	0.92	0.34	0.49	0.76	0.58	0.88	0.76
3	0.95	0.35	0.51	0.76	0.58	**0.89**	0.76
4	0.96	0.38	0.52	0.78	0.60	0.86	0.78
5	**0.97**	0.39	0.53	0.78	**0.62**	0.82	0.79
6	0.96	**0.41**	0.52	**0.80**	**0.62**	0.62	0.78
7	0.94	0.32	0.51	0.77	0.59	0.10	0.76
8	0.95	0.39	**0.54**	0.79	0.61	0.23	0.77
9	0.95	0.40	**0.54**	0.79	**0.62**	0.32	0.78
10	0.95	0.40	**0.54**	**0.80**	**0.62**	0.31	0.81
11	0.96	0.38	0.52	0.78	0.60	0.34	0.80
12	0.95	0.36	0.52	0.77	0.59	0.47	**0.82**
T5 (English); Layers 7–12 are the decoder; std = 0.00 – 0.03.							
1	0.77	0.27	0.39	0.71	0.50	0.26	0.71
2	0.80	0.30	0.43	0.74	0.54	0.38	0.70
3	0.82	0.32	0.45	0.75	0.55	**0.40**	0.73
4	0.84	0.33	**0.46**	**0.76**	0.57	0.37	**0.74**
5	0.87	0.33	0.45	**0.76**	0.57	0.33	0.72
6	**0.86**	**0.35**	0.46	**0.76**	**0.58**	0.28	0.72
7	0.77	0.28	0.41	0.73	0.54	0.24	0.71
8	0.77	0.26	0.44	0.74	0.55	0.27	0.71
9	0.77	0.24	**0.46**	0.75	0.56	0.27	0.71
10	0.74	0.22	0.45	0.75	0.55	0.20	0.72
11	0.70	0.22	0.44	0.73	0.54	0.11	0.72
12	0.68	0.20	0.42	0.73	0.52	0.00	0.72

Table 10: Full results for GPT-2, BART, and T5 over English.

Layer	**NSP**	**Sent. Ord.**	**Discourse Conn.**	**Nuclearity**	**Relation**	**EDU segment.**
Chinese; std = 0.00 – 0.02.						
1	0.30	0.38	0.36	0.53	0.58	0.70
2	0.47	0.38	0.36	0.53	0.58	0.75
3	0.47	0.42	0.36	0.53	0.58	0.79
4	0.83	0.42	0.36	0.53	0.58	0.79
5	0.90	0.44	0.35	0.59	0.60	0.81
6	0.93	0.44	0.39	0.60	0.64	**0.83**
7	0.94	**0.45**	0.42	**0.64**	0.67	**0.83**
8	0.94	0.44	0.43	**0.64**	0.66	**0.83**
9	0.94	0.43	**0.46**	0.63	**0.68**	**0.83**
10	0.96	0.43	0.44	**0.64**	**0.68**	**0.83**
11	0.96	0.42	**0.46**	0.61	0.66	0.81
12	**0.98**	0.40	0.44	0.62	0.66	0.78
German; std = 0.00 – 0.07.						
1	0.43	0.58	0.44	0.48	0.16	0.55
2	0.60	0.58	0.43	0.49	0.18	0.56
3	0.77	0.59	0.43	0.47	0.17	0.67
4	0.76	0.61	0.43	0.44	0.21	0.71
5	0.98	0.63	0.38	0.48	0.26	0.73
6	0.99	**0.65**	0.43	0.54	0.29	0.74
7	**1.00**	**0.65**	0.43	0.58	0.31	**0.76**
8	0.99	0.64	0.43	**0.60**	**0.35**	0.75
9	**1.00**	0.64	0.44	0.59	0.33	0.69
10	0.99	0.64	**0.45**	0.58	**0.35**	0.65
11	**1.00**	0.63	0.43	0.58	0.33	0.58
12	**1.00**	0.63	0.38	0.58	0.33	0.59
Spanish; std = 0.00 – 0.02.						
1	0.39	0.49	—	0.50	0.29	0.43
2	0.55	0.52	—	0.56	0.31	0.50
3	0.56	0.53	—	0.58	0.31	0.52
4	0.96	0.55	—	0.62	0.37	0.57
5	0.98	0.56	—	0.64	0.41	0.59
6	0.99	0.56	—	0.68	0.45	0.62
7	**1.00**	0.57	—	0.68	0.47	0.64
8	**1.00**	**0.58**	—	0.75	0.49	**0.69**
9	**1.00**	**0.58**	—	0.74	0.51	0.66
10	**1.00**	**0.58**	—	**0.77**	**0.56**	0.62
11	**1.00**	0.57	—	**0.77**	0.55	0.59
12	**1.00**	0.56	—	0.76	0.54	0.50

Table 11: Full results for the BERT monolingual models over Chinese, German, and Spanish.

Layer	**NSP**	**Sent.**	**Discourse**	**Nuclearity**	**Relation**	**EDU**	**Cloze**
		Ord.	**Conn.**			**segment.**	**ST.**
BERT-Large (English); std = 0.00 – 0.02.							
1	0.34	0.26	0.29	0.60	0.37	0.23	0.61
2	0.53	0.25	0.33	0.62	0.40	0.28	0.67
3	0.57	0.26	0.32	0.63	0.42	0.29	0.66
4	0.60	0.29	0.32	0.63	0.42	0.40	0.66
5	0.64	0.30	0.35	0.65	0.43	0.43	0.69
6	0.82	0.31	0.37	0.66	0.44	0.45	0.68
7	0.87	0.31	0.39	0.66	0.45	0.44	0.69
8	0.95	0.30	0.39	0.66	0.45	0.44	0.72
9	0.96	0.32	0.40	0.67	0.47	0.43	0.70
10	0.96	0.32	0.40	0.67	0.46	0.49	0.71
11	0.97	0.32	0.39	0.68	0.47	0.53	0.70
12	0.97	0.33	0.47	0.77	0.57	0.58	0.71
13	0.97	0.34	0.52	0.78	0.61	0.56	0.71
14	0.97	0.34	0.57	0.81	0.65	**0.59**	0.73
15	0.97	0.34	0.61	0.82	0.67	0.58	0.75
16	0.97	0.35	0.60	**0.83**	0.67	0.55	0.75
17	0.97	0.34	0.62	0.82	**0.68**	0.53	0.82
18	0.98	0.36	**0.63**	0.82	**0.68**	0.54	0.82
19	**0.99**	**0.37**	**0.63**	0.82	0.67	0.50	0.83
20	**0.99**	0.34	**0.63**	0.81	0.67	0.48	**0.84**
21	**0.99**	0.35	**0.63**	0.81	0.65	0.41	0.83
22	**0.99**	0.35	0.61	0.81	0.65	0.37	**0.84**
23	**0.99**	0.34	0.59	0.80	0.63	0.36	0.82
24	**0.99**	0.33	0.57	0.77	0.58	0.31	0.81

Table 12: Full results of English BERT-large.

D Frozen vs. Fine-tuned BERT Layers

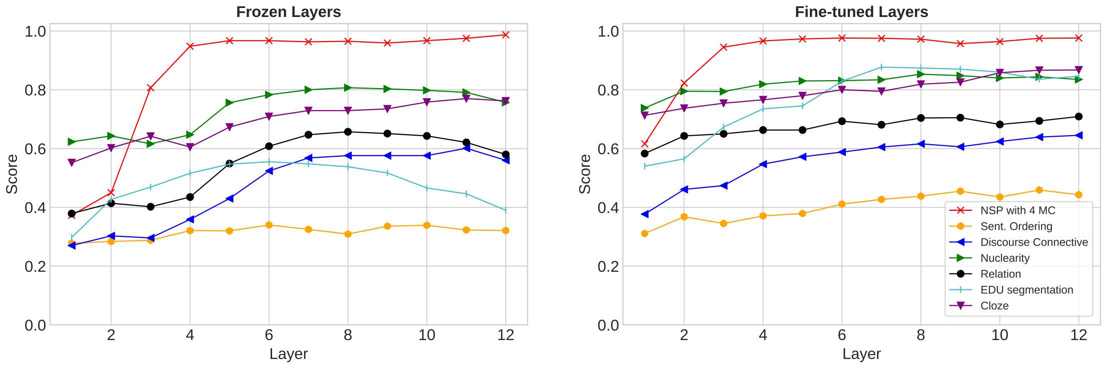


Figure 9: A comparison of BERT with frozen vs. fine-tuned layers.

E Full Results of Models with Average Pooling

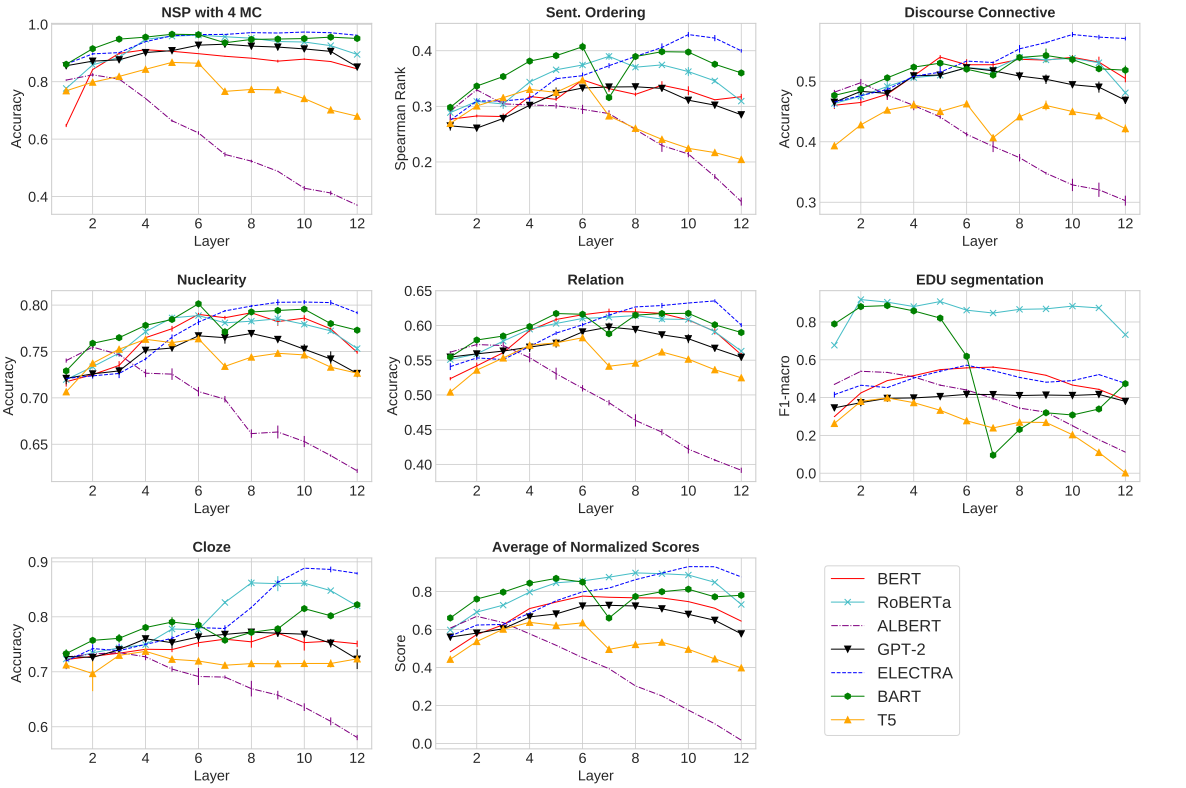


Figure 10: Full results of all models over English with average pooling on all tasks except in EDU segmentation (with the only differences over Figure 2 being for BERT and ALBERT, where we originally used [CLS] embeddings on two-text classification probing tasks).

F [CLS] vs. Average Pooling in English BERT-base Model

Average pooling generally performs worse than [CLS] embeddings in the last layers of BERT.

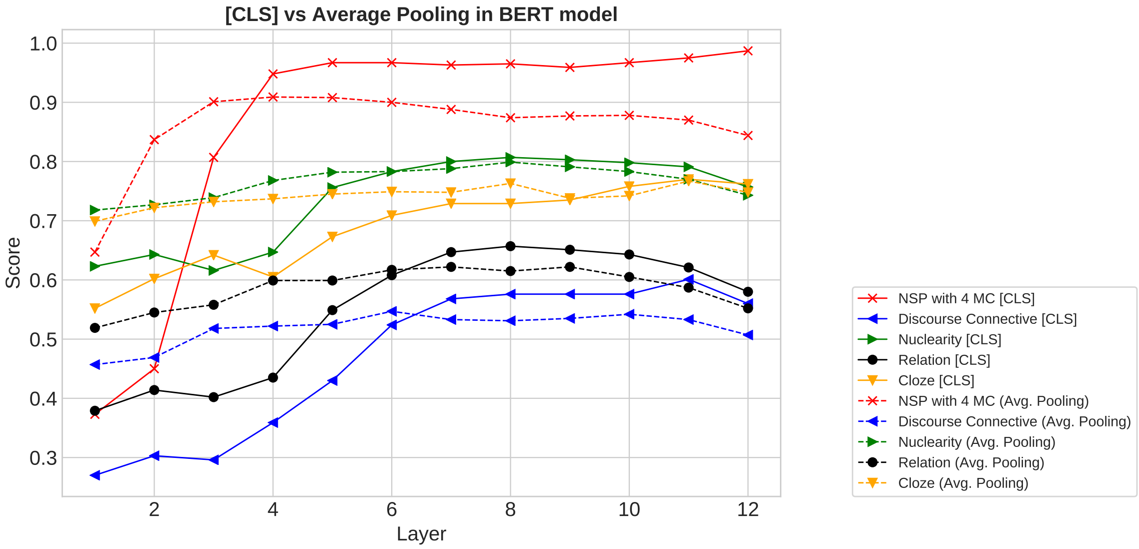


Figure 11: Comparison of [CLS] vs. average pooling embeddings for BERT-base across the five tasks for English. Please note that sentence ordering and EDU segmentation are always performed with average pooling embeddings and sequence labelling at the (sub)word level, respectively.