Evaluating Discourse Cohesion in Pre-trained Language Models

Jie He[†], **Wanqiu Long**[†], and **Deyi Xiong**[‡]
[†] University of Edinburgh, Edinburgh, UK
[‡] College of Intelligence and Computing, Tianjin University, Tianjin, China
j.he@ed.ac.uk, Wanqiu.long@ed.ac.uk, dyxiong@tju.edu.cn

Abstract

Large pre-trained neural models have achieved remarkable success in natural language process (NLP), inspiring a growing body of research analyzing their ability from different aspects. In this paper, we propose a test suite to evaluate the cohesive ability of pre-trained language models. The test suite contains multiple cohesion phenomena between adjacent and non-adjacent sentences. We try to compare different pre-trained language models on these phenomena and analyze the experimental resultshoping more attention can be given to discourse cohesion in the future. The built discourse cohesion test suite will be publicly available at https://github.com/probe2/discourse_cohesion.

1 Introduction

Pre-trained language models have achieved remarkable success in many downstream tasks, including question answering (Wang et al., 2019), reading comprehension (Yang et al., 2019), and machine translation (Imamura and Sumita, 2019), inspiring a growing body of research analyzing their ability from different aspects (Ethayarajh, 2019; Joshi et al., 2019). However, to our best knowledge, there is no existing work to evaluate whether the abilities of these models to identify and generate discourse cohesion.

Cohesion is the foundation of an essay and an important form of showing style and character, and it is a semantic property of a document that represents the degree to which discourse entities are knit throughout the document (Li, 2013; Bhatnagar et al., 2022). Halliday et al. (1976) defined cohesion as "the set of possibilities that exist in the language for making text hang together". Cohesion occurs where the interpretation of some element in the discourse is dependent on that of another. For example, an understanding of the reference of a pronoun (he, she, it, etc.) requires to look back to something that has been said before. Through this cohesion relation, two text clauses or sentences are linked together. Therefore, cohesion plays an important role in discourse.

However, to our best knowledge, existing available resources either only provide annotations for one cohesive phenomenon or mainly focus on lexical cohesion. For example, Bos and Spenader (2011) annotate verbal phrase ellipsis; Martínez et al. (2016) annotate lexical cohesion for both German and English texts. However, neither single cohesion phenomena nor just lexical cohesion can fully interpret the ability of models from the perspective of cohesion.

Considering the above, this work has the following contributions:

- We study discourse cohesion for pre-trained language models, which has been understudied in previous works on representation learning, but is critical to language understanding and generation.
- We propose a test suite of cohesion including both grammatical and lexical cohesion phenomena.
- We conduct a qualitative analysis of different pre-trained language models for their ability for multiple cohesion phenomena from both adjacent and non-adjacent sentences.

2 Related work

Discourse Cohesion Modeling
Some discourse cohesion phenomena have been applied in various NLP tasks. A thorough survey of related work on this is far beyond the scope of this paper. To name just a few, Voita et al. (2019) study repetition and ellipsis in machine translation; Geva et al. (2019) tried to bring the connection between two sentences closer by combining rule-based methods with coreference and conjunction. Similarly, there are also some works dedicated to

Proceedings of 3rd Workshop on Computational Approaches to Discourse, pages 28–34
October 16 – 17, 2022.

cohesion phenomenon	Category	Example	Size
Repetition	adj	he decided to buy a **pair** of khakis. the **pair** he bought fit him perfectly.	200
	non-adj	Jude was very excited about his college graduation **ceremony**. On the way to the arena, he got stuck in traffic. He only had an hour before the **ceremony** started.	73
Synonyms	adj	jill became very **scared**. liam could tell jill was truly **frightened**.	200
	non-adj	She decided not to pursue the **matter** and just keep the service. It was after all only $12. But the **issue** kept bothering her.	64
Ellipsis	adj	But **we** have an interest in **hiring him**; I just don't know **when**.	200
	non-adj	Shawn felt that he could learn **to make the website on his own**. Due to budget he could not pay a web designer. He took many web development classes to learn **how**.	50
Substitution	adj	She wanted those **cookies**. She then decided to take **one**.	200
	non-adj	She began to drink a few **beers**. He had never been a drinker. She encouraged him to drink **one**.	61
Reference	adj	At first he did not like the **classes**. however, over time he began to like **them** a lot.	200
	non-adj	Once there **Jill** marveled at all the beauty. It was dangerous, but exciting. **She** had a wonderful time on her trip to the Amazon.	51
Conjunction	adj	it was also cash only. **therefore** i had to turn around and go home.	200
	non-adj	The couple rented a yurt. It was very small. They did not like being so close. They left the Yurt. They rented a hotel **instead**.	55

Table 1: Examples of cohesion phenomena adopted in our test suite. Repetition and synonyms are lexical cohesion. Non-adj means the cohesion phenomenon is annotated between non-adjacent sentences, while adj refers to cohesion between adjacent sentences.

the study of discourse phenomena. For example, Uryupina et al. (2020) annotated a broad range of anaphoric phenomena in a variety of genres. Pishdad et al. (2020) studied the phenomenon of coherence at both the lexical and document levels. We are the first work to evaluate the performance of the pre-trained language model about multiple discourse cohesion phenomena.

Analysis towards Pre-trained Language Models The boom of pre-trained language models has stimulated plenty of work to probe into the internal working mechanisms and capacities of pre-trained language models (Liu et al., 2019b; Joshi et al., 2019; Lewis et al., 2020). For example, Jawahar et al. (2019) investigate the ability of these pre-trained models from the structure of language; Liu et al. (2019a); Warstadt et al. (2020) analyze those models from syntactic phenomena. Chen et al. (2019) study whether sentence representations from pretrained language models contain contextual information. Meanwhile, Kim et al. (2019) test pre-trained language models for functional words within sentences.

However, although there are resources annotated for individual phenomena separately, there are not so many annotated for several types of devices, so no existing work tries to simultaneously evaluate whether the pre-tained language models are good enough for identifying and generating different multiple cohesion phenomena and to compare and analyze the results.

3 Our Test Suite and its Annotation

3.1 Introduction

Halliday et al. (1976) describe five main types of cohesion in English, which we adopt for our suite: reference, substitution, ellipsis, conjunction and lexical cohesion. Table 1 demonstrates the examples and size for the six cohesion phenomena covered in our test suite. The test suite contains 1554 cohesion examples in total. While cohesive cohesion have in principle noting to do with sentence boundaries (Halliday et al., 1976), we take into account cohesive relations between adjacent sentences/clauses as well as those between non-adjacent sentences. However, due to the data sparsity, there are 354 instances in total between non-adjacent sentences, while each phenomenon has 200 instances between adjacent sentences.

The cohesion examples for six cohesion phenomena in this test suite were all drawn from the ROC stories corpus (Mostafazadeh et al., 2016). There are 50k five-sentence commonsense stories in this corpus. This corpus is a high quality collection of everyday life stories, which captures a rich set of relations between daily events.

3.2 Lexical Cohesion

Lexical cohesion arises from the semantic relationship between words, as the chains of related words can generate the continuity of lexical meaning. Two typical ways of achieving this kind of cohesion is repetition and synonyms.

Repetition: Repetition means the repeating of certain words or phrases. The task is to study the relationship between repeated words from two sentences, while our dataset for this phenomenon is on the nouns repetition.

Synonyms: As for synonyms, it means there are related words that having the same connotations, implications, or reference in two sentences. Therefore, the task is to observe whether the synonyms from two sentences are magnets for each other in the models. In our test suite, the sentence pairs for this phenomenon include nouns indicating synonyms.

3.3 Grammatical Coehsion

Our grammatical cohesion tasks investigate whether the models have the ability to identify the anaphoric relationship between entities or how the sentences are connected with each other.

Reference: Reference is a relationship between objects in which one object designates, or acts as a means by which to connect to or link to, another object.

Substitution: Substitution generally occurs when one item within a text or discourse is replaced by another. The examples for this phenomenon are mainly represented by the substitution of nouns by using "one". For instance, "this house is old. I will buy a new one".

Ellipis: Ellipsis means the omission of one or more words that are obviously understood but that must be supplied to make a construction grammatically complete. For this part of the data, we use the sluice ellipsis dataset (Anand and McCloskey, 2015), which studies the omission after wh-words.

Conjunction: Unlike other grammatical cohesion phenomena, conjunction expresses a logical semantic relationship between two sentences rather than between words or structures. According to Halliday et al. (1976), conjunction can be divided into 4 categories: additive, adversative, causal, and temporal. In our test set, we covered these 4 categories.

Markers: Although without discourse markers, the meaning of the sentences would not be affected, they enable the connection between sentences to stick together.

3.4 Annotation

To construct the test suite, we hired 2 fluent English speakers to manually annotate data.

Since cohesion is something available in the surface structure, it is relatively easy to identify. Therefore, we were able to filter a great number of sentences without cohesion by using the "cohesive devices" and WordNet (Fellbaum, 2000). Cohesive devices are words or phrases used to connect ideas between different parts of text. From Table 1, we can see "one", "when", "how", "therefore", etc. as "cohesive devices". WordNet was used to identify synonyms.

However, the automatic filtering is just the first step. Human annotation is necessary since most automatically selected sentences have no cohesion. Before manual annotation, our annotation guidance and requirements were explained in detail to the annotators:

- The annotators are required to observe whether the sentence has corresponding phenomena. For example, the repetition phenomenon requires the nouns that refer to the same thing to appear twice in the sentence. The phenomenon of ellipsis requires ellipsis hint words (wh-words here) to appear in the sentence.
- After identifying whether certain cohesion phenomenon is shown, the annotators needs to mark the two elements that convey cohesion. If the two elements that convey cohesion cannot be marked, the sentence would not be used.

To ensure annotation consistency, we compute the Kappa value and agreement rate between two annotators for agreement study. Before annotation, we randomly selected 500 examples as samples for pre-annotation, then two annotators labelled the text in terms of our annotation guidelines respectively. Finally, we got the average IAA and Cohen's kappa value for the two annotators' annotation, which is 91.3% and 80.6%.

4 Experiments

4.1 Models

We chose the pre-trained language model BERT (Devlin et al., 2019), BART (Lewis et al., 2020) and RoBERTa (Liu et al., 2019b) as our evaluation

Model	Repetition adj	Repetition non-adj	Synonym adj	Synonym non-adj	Reference adj	Reference non-adj	Substitution adj	Substitution non-adj	ellipsis adj	ellipsis non-adj	conjunction adj	conjunction non-adj
BERT-base	0.690	0.493	0.240	0.391	0.830	0.510	0.365	0.262	0.421	0.180	0.235	0.364
BERT-large	0.730	0.644	0.270	0.469	0.850	0.608	0.470	0.328	0.455	0.280	0.340	0.455
BART-base	0.725	0.795	0.215	0.422	0.675	0.490	0.375	0.180	0.302	0.34	0.135	0.018
BART-large	0.710	0.740	0.250	0.500	0.715	0.627	0.390	0.230	0.302	0.260	0.100	0.145
RoBERTa-base	0.780	0.712	0.325	0.469	0.790	0.804	0.545	0.377	0.624	0.540	0.395	**0.673**
RoBERTa-large	**0.815**	**0.836**	**0.430**	**0.594**	**0.855**	**0.863**	**0.665**	**0.393**	**0.678**	**0.600**	**0.485**	0.655
HUMAN	0.86	0.72	0.83	0.915	0.952	0.810	0.876	0.780	0.865	0.820	0.925	0.840

Table 2: Accuracy of the masked-word-prediction

models. The pretraining task of BART involves randomly shuffling the order of the original sentences and a novel in-filling scheme, where spans of text are replaced with a single mask token. While BERT and RoBERTa mainly differ in their training set size, BERT and BART is different in their training methods and model architectures.

4.2 Cohesion Evaluation

We would like to investigate whether the pretrained language models capture enough knowledge related to cohesion. We evaluated model performance via the prediction of masked words. A masked-word-prediction head (either fine-tuned or not) produces a probability distribution over its whole vocabulary via a softmax layer. We consider hit@1, namely the word filled with the highest probability when evaluating. If the hit@1 generated is able to link two clauses or sentences together, we think the model show the ability of identifying and generating cohesion. For example, in this example, "he decided to buy a pair of khakis. The [MASK] he bought fit him perfectly." , "pair" would be expected to be filled when considering repetition.

Besides, to investigate whether the models utilize the context, we compare the probability of generating the target word with and without the previous sentences/clauses on the sub-testset of cohesion between adjacent sentences. In the example, "he decided to buy a pair of khakis. The [MASK] he bought fit him perfectly.", we compare the probability of generating the target word "pair" with and without the span of "he decided to buy a pair of khakis". Finally, we got average probability of the target words for the six cohesion phenomena in both situations.

4.3 Results

Table 2 displays the result of our evaluation task. Firstly, we can see that RoBERTa is the best model in terms of their performance on all cohesion phenomena. BART is inferior to BERT in many phenomena such as synonyms, reference, subsitution, ellipsis. This indicates that the pre-training task of BART may not be very helpful for understanding discourse cohesion phenomena.

From table 2, we can see that conjunction, substitution, synonym and ellipsis are more complicated cohesion types, because the pre-trained language models are not good at them, compared with other cohesion phenomena. With regard to synonyms, it requires that the models not only can identify the cohesion but also have awareness of paraphrasing, which makes it difficult for the models. Looking at the data, we found that the RoBERTa tends to repeat the same word instead of generating another similar word to express the same meaning, even when it notices there is cohesion between the word that should be covered and the corresponding word. In other words, if the models fail to find other cohesive ways, they would try to repeat the words they identify to convey cohesion.

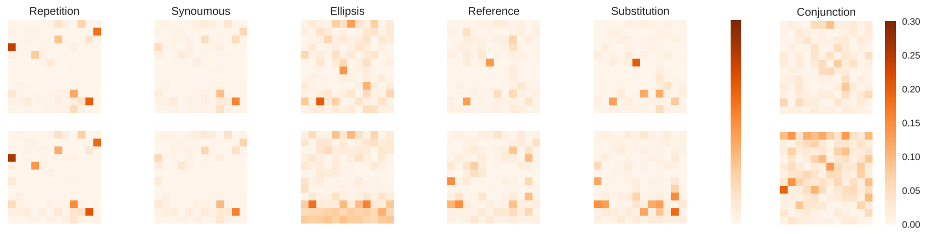
Moreover, model performance on cohesion phenomena between adjacent sentences and non-adjacent sentences can be compared by looking at the Table 2. The models perform better for the cohesion phenomena between non-adjacent sentences instead of adjacent sentences except for substitution. It might be because additional sentences between the two cohesive elements provide context for the models to identify those cohesion phenomena.

5 The probability of generating the target word

Table 3 gives us the information about the probability of generating the target word with and without providing the previous sentences/clauses. From the results of table 3, we can see without the previous sentence/clause, the possibilities of generating the target word for all cohesion phenomena are greatly

Model	Repetition		Synonym		Reference		Substitution		ellipsis		conjunction	
	w/o-C	w/-C	w/o-C	w/-C	w/o-C	w/-C	w/o-C	w/-C	w/o-C	W-C	w/o-C	w/-C
BERT-base	0.085	0.510	0.083	0.173	**0.262**	0.664	0.061	0.266	0.257	0.338	0.050	0.082
BERT-large	0.116	0.557	0.100	0.209	0.238	**0.737**	0.060	0.363	**0.260**	0.399	**0.061**	**0.098**
BART-base	0.047	0.392	0.050	0.105	0.052	0.279	0.023	0.172	0.103	0.207	0.002	0.003
BART-large	0.045	0.309	0.061	0.128	0.067	0.337	0.031	0.209	0.127	0.233	0.002	0.003
RoBERTa-base	0.109	0.585	0.106	0.223	0.155	0.507	0.062	0.407	0.221	0.457	0.009	0.031
RoBERTa-large	**0.144**	**0.662**	**0.114**	**0.268**	0.175	0.652	**0.079**	**0.515**	0.257	0.52	0.01	0.075

Table 3: Probability of the target word with and without prior context.



(a) (b)

Figure 1: Attention heatmaps for 7 types of discourse phenomena.

decreased. Therefore, there is strong cohesion between the target word in the second sentence and the corresponding word in the first sentence. However, the context provided by the first sentence have little positive impacts on BART for these cohesion phenomena, compared with other models.

6 Internal Analysis of BERT for Cohesion Phenomena

For these 7 kinds of cohesion phenomena, we got some fine-grained information from the attention heatmap. The upper part of Figure 1(a) indicates the attention between the words of sentence/clause one and the words of the second sentence/clause two, while the below of Figure 1(a) demonstrates the attention between the words of sentence two and sentence one. We note that repetition and synonym have great attention in both directions, with almost equivalent attention. This explains why the models are better at identifying these two cohesion phenomena. What's more, the attention mainly gather on the deeper layers, which might reflect the deeper layers of BERT capture more complex semantic features.

In Figure 1(b), the upper part represents the attention between the first sentence and the conjunction word/discourse marker, whereas the below represents the attention between the second sentence and the conjunction word or discourse marker. The attention heatmap shows that much more attention can be seen between sentence two and the words, which means that the conjunction word or discourse marker is more closely related to the second sentence. However, it can be observed that the maximum attention of all head value for these two phenomena does not exceed 0.3, thus illustrating the poor performance of the pre-trained language models on these two phenomena is largely due to insufficient attention between the conjunction words or discourse markers and the sentences.

7 Conclusion

We have created a benchmark test suite to evaluate the ability of pre-trained language models on seven discourse cohesion phenomena. And we consider the cohesion phenomena between adjacent sentences/clauses and non-adjacent sentences. Moreover, we conduct analysis on the results of different pre-trained language models for six discourse cohesion phenomena. In the future, we would like to know the capability of language models in terms of global cohesion.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. The corresponding author is Deyi Xiong (dyxiong@tju.edu.cn).

References

Pranav Anand and Jim McCloskey. 2015. Annotating the implicit content of sluices. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 178–187, Denver, Colorado, USA. Association for Computational Linguistics.

Vasudha Bhatnagar, Swagata Duari, and S. K. Gupta. 2022. Quantitative discourse cohesion analysis of scientific scholarly texts using multilayer networks.

Johan Bos and J. Spenader. 2011. An annotated corpus for the analysis of vp ellipsis. *Language Resources and Evaluation*, 45:463–494.

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 (EMNLP-IJCNLP)*, pages 649–662, Hong Kong, China. 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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

C. Fellbaum. 2000. Wordnet : an electronic lexical database. *Language*, 76:706.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. DiscoFuse: A large-scale dataset for discourse-based sentence fusion. 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 3443–3455, Minneapolis, Minnesota. Association for Computational Linguistics.

M.A.K. Halliday, R. Hasan, R.H. Halliday, Pearson Longman, and R. Quirk. 1976. *Cohesion in English*. A Longman paperback. Longman.

Kenji Imamura and Eiichiro Sumita. 2019. Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, Hong Kong. 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.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. 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 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Junxin Li. 2013. The application and significance of discourse cohesion and analysis in practical teaching of foreign language. *Theory and Practice in Language Studies*, 3:1393–1398.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. 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. 2019b. Roberta: A robustly optimized bert pretraining approach.

José Manuel Martínez Martínez, Ekaterina Lapshinova-Koltunski, and K. Kunz. 2016. Annotation of lexical cohesion in english and german: Automatic and manual procedures. In *KONVENS*.

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.

Leila Pishdad, Federico Fancellu, Ran Zhang, and Afsaneh Fazly. 2020. How coherent are neural models of coherence? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6126–6138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1):95–128.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. 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 5878–5882, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2346–2357, Florence, Italy. Association for Computational Linguistics.