CLiMP: A Benchmark for Chinese Language Model Evaluation

Beilei Xiang,[1] **Changbing Yang,**[1] **Yu Li,**[1] **Alex Warstadt**[2] and **Katharina Kann**[1]

[1]University of Colorado Boulder, [2]New York University

{beilei.xiang, changbing.yang, yuli9309}@colorado.edu
warstadt@nyu.edu
katharina.kann@colorado.edu

Abstract

Linguistically informed analyses of language models (LMs) contribute to the understanding and improvement of these models. Here, we introduce the corpus of Chinese linguistic minimal pairs (CLiMP), which can be used to investigate what knowledge Chinese LMs acquire. CLiMP consists of sets of 1,000 minimal pairs (MPs) for 16 syntactic contrasts in Mandarin, covering 9 major Mandarin linguistic phenomena. The MPs are semi-automatically generated, and human agreement with the labels in CLiMP is 95.8%. We evaluate 11 different LMs on CLiMP, covering *n*-grams, LSTMs, and Chinese BERT. We find that classifier–noun agreement and verb complement selection are the phenomena that models generally perform best at. However, models struggle the most with the *bǎ* construction, binding, and filler-gap dependencies. Overall, Chinese BERT achieves an 81.8% average accuracy, while the performances of LSTMs and 5-grams are only moderately above chance level.

1 Introduction

Language models (LMs) are crucial parts of natural language processing (NLP) systems for a large variety of tasks, including summarization, machine translation, and dialog generation. More recently, they have become popular in the form of pretrained models,[1] which are then fine-tuned on downstream tasks and often obtain state-of-the-art performance (Peters et al., 2018; Devlin et al., 2019; Conneau et al., 2020). However, which linguistic phenomena language models can or cannot learn is still poorly understood for many languages.

Resources for the syntactic evaluation of LMs, such as BLiMP (Warstadt et al., 2020) have focused mainly on English, and non-English resources currently only cover a small set of phenomena (Mueller et al., 2020; Gulordava et al., 2018; Ravfogel et al., 2018). In order to spur the analysis and subsequent improvement of LMs in Chinese, we introduce the corpus of Chinese linguistic minimal pairs (CLiMP), which can be used to evaluate LMs' knowledge of Chinese grammar.

CLiMP consists of 16 individual datasets that are semi-automatically generated from grammar templates. Each set—or *paradigm*—contains 1,000 minimal pairs (MPs). Together, they cover 9 core linguistic phenomena in Chinese. Human agreement on this corpus is 95.8%, confirming that CLiMP represents robust contrasts in Chinese grammar. High performance on CLiMP thus implies high correlation with human acceptability judgments across these phenomena.

We use CLiMP to study Chinese BERT (Devlin et al., 2019),[2] 6 LSTM (Hochreiter and Schmidhuber, 1997) LMs, and 4 5-gram LMs. We evaluate for each MP whether the LM assigns a higher probability to the grammatical or the ungrammatical sentence. Our results show that Chinese BERT is closest to human performance, achieving an 81.8% accuracy on average over all phenomena, while the performances of LSTMs and 5-grams, regardless of the training data size, are only moderately above chance level. Classifier–noun agreement and verb complement selection are the phenomena that models generally perform best at, suggesting that Chinese LMs are better at acquiring knowledge of local selectional restrictions. The *bǎ* construction, binding, and filler-gap dependencies are the phenomena models have the most difficulties with. This indicates that they struggle to learn hierarchical syntax and to identify long-distance dependencies.

---

[1]Throughout this paper, we adopt a broad definition of LMs, which includes language representation models which have been trained on a masked language modeling objective.

[2]https://github.com/google-research/bert/blob/master/multilingual.md

2 Related Work

2.1 Language Models

LMs assign probabilities to sequences of words (Jurafsky and Martin, 2009). Recently, they have become commonly used as pretrained models, which can be fine-tuned for downstream NLP tasks (Peters et al., 2018; Devlin et al., 2019; Conneau et al., 2020). Strictly speaking, LMs compute the probabilities of words based only on past context. BERT (Devlin et al., 2019), however, is trained using a masked language modeling objective: it predicts words based on past and future tokens. Wang and Cho (2019) show that BERT is a Markov random field language model that can assign sentences a pseudo-log-likelihood score, which is computed by summing the conditional log probabilities of all tokens in the sentence, as well as generate text. Shin et al. (2019) and Salazar et al. (2020) apply pseudo-log-likelihood scores to sentence ranking and LM evaluation.

2.2 Evaluation of Linguistic Knowledge

Numerous methods exist for probing syntactic knowledge of neural network models in English (Hewitt and Manning, 2019; Tenney et al., 2019), and a growing body of work evaluates the syntactic knowledge of neural models by testing whether they can judge the grammatical acceptability of sentences. One common version of this task uses MPs to evaluate LMs' linguistic knowledge (Linzen et al., 2016; Marvin and Linzen, 2018; Warstadt et al., 2020; Wilcox et al., 2018).

A MP is a pair of sentences that only differ in acceptability due to a single edit, as in (1) and (2). Native speakers can be asked to choose which sentence in each pair sounds more grammatical. Semi-automatically generating MPs can yield a larger set of controlled sentences, providing sufficient data for model evaluation (Linzen et al., 2016; Marvin and Linzen, 2018; Ettinger et al., 2018).

(1)　王鑫　 把 自行车　 扔　 了。
　　 Wángxīn bǎ zìxíngchē rēng le
　　 SUBJ. BA. OBJ.　 V. PST.
　　 "Xin Wang threw away a bike."

(2)　王鑫　 被 自行车　 扔　 了。
　　 Wángxīn bèi zìxíngchē rēng le
　　 SUBJ. PASS. OBJ.　 V. PST.
　　 "Xin Wang was thrown away by a bike."

It is possible to model acceptability in a totally unsupervised way using LMs. The model assigns a probability to each sentence in a MP, and the one with the higher score is predicted as correct, and the model's predictions can be evaluated against human judgments (Marvin and Linzen, 2018; Warstadt et al., 2020). Supervised approaches are also possible (Warstadt et al., 2019), but can be less informative on LMs' linguistic knowledge acquisition due to the bias introduced by training on acceptability judgment labels.

Some prior work evaluates the linguistic knowledge of different non-English models (Ravfogel et al., 2018; Gulordava et al., 2018; Mueller et al., 2020). However, these efforts focus mainly on subject-verb agreement, which is absent in Chinese, and the knowledge of Chinese LMs has not yet been explicitly studied.

Finally, the linguistic abilities of English BERT have been investigated in a a lot of prior work, e.g., Clark et al. (2019); Vig (2019); Hewitt and Manning (2019). We refer the reader to Rogers et al. (2021) for an overview.

3 CLiMP

Our main contribution is CLiMP, a corpus of Chinese MPs designed to evaluate Chinese LMs. CLiMP consists of 1,000 MPs for each of 16 grammatical contrasts, covering 9 major Chinese linguistic phenomena. Example MPs for each phenomenon are shown in Table 1.

3.1 Data Generation

We generate data from grammar templates for every paradigm we incorporate. Our templates set lexical, syntactic, and semantic constraints for each paradigm, aiming at building robust contrasts and keeping the sentence length the same within each MP. We then build an annotated vocabulary, and generate sentences by sampling words from it. (1) and (2) show an MP together with the template[3] used to create it.

3.2 Vocabulary

We translate Warstadt et al.'s (2020) English vocabulary, containing 3,000 English words with morphological, syntactical, and semantic annotations. We add words and features specific to Chinese linguistic phenomena to our vocabulary, including classifiers, verb complements, action verbs, and

---

[3]The template example is only for demonstrative purposes. More information is encoded for the actual data generation.

coverbs. Our final vocabulary contains 3,456 words and 84 features.

We show the frequency of words in CLiMP's vocabulary in the Chinese Internet Corpus[4] in Figure 1. 1,055 of the words in CLiMP are within the 5,000 most frequent words in the Chinese Internet Corpus.
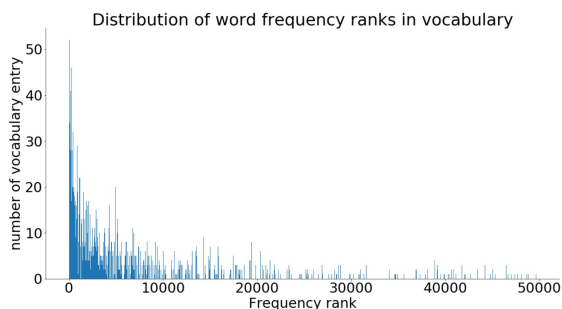

Figure 1: Comparison of word frequencies in CLiMP and the Chinese Internet Corpus.

3.3 Linguistic Phenomena

CLiMP covers 9 major linguistic phenomena in Mandarin Chinese, cf. Table 1. They are picked from a comprehensive Chinese grammar book by Po-Ching and Rimmington (2015). Following Po-Ching and Rimmington's discussion, we now explain the phenomena not present in English. The *bǎ* **construction** is an SOV construction involving the particle *bǎ*, which precedes the object and moves the object to a position before the main verb. It is only grammatical with a subset of transitive verbs. **Coverbs** are verb-like items that precede the main verb in a serial verb construction. They almost invariably have to be used in conjunction with other verbs in a sentence. They share some properties with prepositions, but are not syntactically interchangeable with them. **Classifiers** obligatorily appear with nouns when those are modified by numerals or adjectives. Mandarin has dozens of classifiers, and nouns select the classifier they combine with. **Verb complements** follow a verb, often expressing a result or manner of an event. Not all verbs can be used with all complements, making certain combinations ungrammatical. **NP head finality** is present in Mandarin noun phrases. The relative clause precedes noun phrases.

3.4 Data Validation

To verify whether the MPs in our dataset show clear contrasts, we conduct two rounds of human valida-

tion with 22 annotators. They are all native speakers of Chinese, 14 females and 8 males, whose ages range from 20 to 48. All of them have at least a high school degree.

In our first human validation, each human annotator is assigned a subset (100 MPs) of a paradigm. We let them perform the same forced-choice task as our models: decide for each MP which sentence seems more acceptable. We discard one paradigm, the coverb-direction paradigm, after this validation, because its human validation accuracy is below 85%. The average human agreement for the remaining paradigms is 95.8%.

In the second human validation, we sample 15 MPs from each of the remaining paradigm, resulting in a dataset consisting of 240 MPs. 16 annotators complete the same forced-choice task on this dataset. We count a MP as valid if more than half of the annotators agree with its label. The human agreement on this dataset is 97.1%, showing that our data creation results in valid examples.

3.5 Comparison with BLiMP

BLiMP consists of 67 datasets, each containing 1,000 MPs and organized by phenomenon into 12 categories. CLiMP only contains 16 datasets due to the less inflectional nature of Mandarin Chinese. 3 phenomena are covered by both corpora: anaphor agreement, binding, and filler-gap. The human agreement for these three phenomena in BLiMP is 97.5%, 87.3%, and 86.9%, respectively. The corresponding accuracies in CLiMP are 94.5%, 99%, and 100%, respectively. The overall human agreement for BLiMP is 88.6%, which is 7.2% lower than for CLiMP.

4 Models and Methods

We use accuracy for evaluation. A MP in CLiMP is classified correctly if a LM assigns a higher probability to the grammatical sentence than to the ungrammatical one. We evaluate statistical and neural LMs, including masked LMs. Corpora which contain 0.4M, 2M, and 21.5M sentences are used for further exploration. We also investigate the effect of different tokenizations.[5]

Chinese BERT BERT (Devlin et al., 2019) is a transformer-based neural model (Vaswani et al., 2017). Here, we evaluate Chinese BERT.[6] This

---

[4]http://corpus.leeds.ac.uk/frqc/internet-zh.num

[5]We use character tokenization and word tokenization (https://github.com/fxsjy/jieba).

[6]https://github.com/google-research/bert/blob/master/multilingual.md

Phenomenon	N	Acceptable Example	Unacceptable Example
Anaphor agreement	1	王玉珍　震惊-了　她自己。 Jane.F　shock-PST　_herself_. _'Jane shocked herself.'_	王玉珍　震惊-了　他自己。 Jane.F　shock-PST　_himself_. _'Jane shocked himself.'_
Binding	1	杨颖　治疗 吴宇涛 之后 佩服-过　她自己。 Yang.F　cure　Wu.M　after admire-PST _herself_ _'Yang admired herself after she cured Wu.'_	杨颖　治疗 吴宇涛 之后 佩服-过　他自己。 Yang.F　cure　Wu.M　after admire-PST　_himself_ _'Yang admired himself after she cured Wu.'_
_bǎ_ construction	1	王鑫　　把 自行车　扔　了。 Wong.M　_BA_　bike throw PST _'Wong threw away the bike.'_	王鑫　　被 自行车　扔　了。 Wong.M　_PASS_　bike throw PST _'Wong was thrown away by the bike.'_
Coverb	3	李文清　乘 卡车　到达-了 咖啡店。 Lee.M　_ride_ truck　arrive-PST coffee shop _'Lee went to the coffee shop by truck.'_	李文清　于卡车　到达-了 咖啡店。 Lee.M　_at_ truck　arrive-PST coffee shop _'Lee went to the coffee shop at truck.'_
NP head finality	1	王梦　　正在 卖 张红梅 清洗-过-的　推车。 Wong.F PROG sell May.F clean-PRF-ADJ trolley _'Wong is selling the trolley that Mel has cleaned.'_	王梦　　正在 卖 推车　张红梅 清洗-过-的。 Wong.F PROG sell trolley May.F clean-PRF-ADJ _'Wong is selling the trolley that Mel has cleaned.'_
Classifier	2	张杰　正在 穿过 一　　家　　艺术画廊。 Jay.M PROG pass one _CL:INSTITUTION_ art gallery _'Jay is passing through an art gallery.'_	张杰 正在 穿过 一　段　　艺术画廊。 Jay.M PROG pass one _CL:LENGTH_ art gallery _'Jay is passing through an art gallery.'_
Filler gap	1	图书馆，　我 开车 去-过 这个地方。 The library,　I drive to-PRF _this place_ _'The library, I have driven to this place.'_	图书馆，　我 开车 去-过　博物馆。 The library,　I drive to-PRF _the museum_ _'The library, I have driven to the museum.'_
Passive	1	这些　患者　被　转移-了。 These patient PASS _transfer_-PST _'These patients were transferred.'_	这些　患者　被　下降-了。 These patient PASS _fall_-PST _'These patients were fell.'_
Verb complement	5	王慧　　的 文章 吓　坏　了 包曼玉。 Wong.F POSS article frighten _badly_ PST　Bao.F. _'Wong's article frightened Bao badly.'_	王慧　　的 文章 吓　开　了 包曼玉。 Wong.F POSS article frighten _openly_ PST　Bao.F. _'Wong's article frightened Bao openly.'_

Table 1: Nine Chinese linguistic phenomena covered by CLiMP with acceptable and unacceptable sentence examples. Minimal differences are underlined. The second line of each example shows a gloss, the third line is an English translation. N represents how many paradigms (each with 1,000 examples) are within each phenomena.

model has 12 layers, 768 hidden units, 12 attention heads, and 110M parameters. The training dataset contains 25M sentences. We assign probabilities to sentences with this model by masking the words in a sentence one by one, computing the probability of each masked word, and, finally, multiplying the probabilities of all words (Wang and Cho, 2019; Salazar et al., 2020).[7]

LSTM LMs We further evaluate 6 LSTM (Hochreiter and Schmidhuber, 1997) LMs. These model have 2 layers, 200 hidden units, and 2 attention heads. We train them using Pytorch's word language model code[8] on 3 differently-sized Chinese Wikipedia corpora: 0.4M, 2M, and 21.5M sentences. We further compare word-level and character-level models (cf. Table 2). For evaluation, we employ code adapted by Warstadt et al. (2020) from Gulordava et al. (2018).[9]

_n_-gram LMs Finally, we experiment with 4 different 5-gram LMs, which have been trained on 0.4M and 2M sentences from Chinese Wikipedia. For each corpus size, we train one word-based and one character-based LM. Those models are imple-

mented using KenLM.[10]

5 Results

All results are shown in Table 2.

Phenomenon-specific Results Our LMs perform best on classifier–noun agreement and verb complement selection: Chinese BERT's accuracy is only 6.8% and, respectively, 3% lower than that of humans on these two phenomena. LSTMs and 5-grams remain around 30% behind humans, but still perform better on these phenomena than on others in CLiMP. This indicates that Chinese LMs acquire local selection knowledge better than the linguistic knowledge needed to master other phenomena.

Our LMs stuggle most with the _bǎ_ construction, binding, and filler-gap dependencies. All models perform close to chance level for binding, suggesting that they lack the hierarchical knowledge necessary to correctly resolve the structural relationship between a reflexive and its binder. Similarly, most models perform near chance on filler-gap dependencies. This suggests that they do not robustly represent long-distance dependencies.[11]

Model	Overall	Clsfr.	V.Cp.	Hd.Fi.	The ba.	Coverb	Ana.Agr.	Pass.	Bind.	Fi.Gap
Human	*95.8*	*99.7*	*96.0*	*100.0*	*85.0*	*92.5*	*94.5*	*91.0*	*99.0*	*100.0*
Chinese BERT	81.8	92.9	**93.0**	53.1	69.0	87.9	86.2	67.7	50.8	62.4
LSTM-21.5M-word	62.8	75.7	74.0	**81.4**	10.0	47.0	63.1	68.4	50.1	41.5
LSTM-21.5M-char	60.7	56.1	64.9	**89.1**	32.1	43.2	57.0	67.9	50.0	68.8
LSTM-2M-word	66.0	**77.8**	73.8	75.0	48.4	43.4	67.0	68.0	50.0	59.2
LSTM-2M-char	60.4	68.4	68.1	**86.3**	29.0	28.5	68.1	68.4	50.1	61.9
LSTM-2M-word	60.6	69.9	65.4	70.3	41.1	38.8	66.3	**72.7**	50.0	55.2
LSTM-2M-char	63.2	68.9	69.7	**83.9**	25.0	45.6	67.7	74.3	50.0	64.4
5-gram-2M-word	59.0	70.1	71	55.2	15.6	39.2	67.7	**72.0**	49.6	40.0
5-gram-2M-char	65.7	70.6	**78.8**	68.3	30.6	53.9	65.8	64.8	51.6	57.3
5-gram-0.4M-word	55.9	66.4	69.5	46.3	6.0	37.0	69.1	**77.8**	49.1	25.2
5-gram–0.4M-char	60.0	**71.5**	65.4	70.5	19.3	46.5	68.8	68.7	50.2	48.4

Table 2: Percentage accuracy of all humans and models on CLiMP. Random guessing yields an accuracy of 50%. Bold numbers indicate the phenomenon each model is best at. Numbers in model names (21.5, 2, 0.4) refer to the number of sentences in the training corpus.

On the head-final construction, Chinese BERT performs surprisingly poorly as compared to the other models: only 53.1% accuracy as compared to an average accuracy of 81% by the LSTMs. The coverb construction, in contrast, is easy for Chinese BERT: it achieves 87.9% accuracy, while the highest accuracy among all other models is 47%.

Model-specific Results Comparing across models, Chinese BERT achieves by far the highest overall accuracy with 81.8%. Our different LSTMs all perform worse, but obtain surprisingly similar scores: from 60.4% to 66.0%. The performances of our *5*-grams range from 55.9% to 65.7%. Keeping tokenization and corpus size constant, three out of four *5*-grams are outperformed by LSTMs. Thus, we overall find that neural models have advantages as compared to statistical models.

Comparing among the LSTMs, we find similarly to Hu et al. (2020) that the corpus size does not have much influence on the overall performance, with the caveat that these models perform close to chance. In contrast, a larger corpus size does result in a better performance in *5*-grams. We also compare the effect of different tokenizations: Character-based *5*-grams demonstrate better performance than word-based ones. For LSTMs, however, using characters only results in a better performance for our smallest corpus size (0.4M).

Compared to English LMs (Warstadt et al., 2020), the human–model gap is much bigger for Chinese models. While neither models nor datasets are directly comparable between our and previous work, this still suggests that more analyses and developments are needed for non-English models.

---

tion more common in speech, and less likely to appear in the training corpora.

6 Conclusion

We introduced CLiMP, a suite of diagnostic test sets aimed at evaluating which syntactic phenomena Chinese LMs learn, and used it to evaluate 11 different models. All LMs appeared to have learned local selectional restrictions, but struggled with argument structure alternations, hierarchical structure, and long-distance dependencies. Chinese BERT performed best on CLiMP overall. However, it obtained a 14% lower accuracy than humans, suggesting there is still much room for improvement. We hope that CLiMP will serve as a linguistically informed resource for benchmarking and analyzing future progress on Chinese LMs. CLiMP is available at https://nala-cub.github.io/resources.

7 Acknowledgments

References

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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.

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.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

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

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Yip Po-Ching and Don Rimmington. 2015. *Chinese: A comprehensive grammar*. Routledge.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can LSTM learn to capture agreement? the case of basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107, Brussels, Belgium. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8(0):842–866.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. 2019. Effective sentence scoring method using BERT for speech recognition. In *Asian Conference on Machine Learning*, pages 1081–1093.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Monananey, Wei Peng, Sheng-Fu Wang, and Samuel Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8(0):377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.