Probing Relational Knowledge in Language Models via Word Analogies

Kiamehr Rezaee and **Jose Camacho-Collados**
Cardiff NLP, School of Computer Science and Informatics
Cardiff University, United Kingdom
{RezaeeK, CamachoColladosJ}@cardiff.ac.uk

Abstract

Understanding relational knowledge plays an integral part in natural language understanding. When it comes to pre-trained language models (PLMs), prior work has been focusing on probing relational knowledge by filling the blanks in pre-defined prompts such as "The capital of France is —". However, these probes may be affected by the co-occurrence of target relation words and entities (e.g. "capital", "France" and "Paris") in the pre-training corpus. In this work, we extend these probing methodologies leveraging analogical proportions as a proxy to probe relational knowledge in transformer-based PLMs without directly presenting the desired relation. In particular, we analysed the ability of PLMs to understand (1) the directionality of a given relation (e.g. *Paris-France* is not the same as *France-Paris*); (2) the ability to distinguish types on a given relation (both France and Japan are countries); and (3) the relation itself (Paris is the capital of France, but not Rome). Our results show how PLMs are extremely accurate at (1) and (2), but have room for improvement for (3). To better understand the reasons behind this behaviour and the types of mistake made by PLMs, we provide an extended quantitative analysis.

1 Introduction

A major area of research in NLP in the past years has been devoted to probing pre-trained language models (PLMs) to measure the extent of which the relational/factual knowledge is captured by their representations (Bouraoui et al., 2020; Jiang et al., 2020; Wallat et al., 2020). Seeking insight into the hidden representational space of PLMs, recent studies have been exploiting the word prediction capabilities of language models in a cloze-style fill-the-blank configuration as a more direct method to probe factual knowledge in PLMs (Petroni et al., 2019; Wallat et al., 2020). For example, in order to query for the capital of Paris, one can use the PLM to fill the blank in a prompt such as 'The capital of Paris is —' by predicting the most probable word as a response.

First, while there have been studies that attempted to find automatic templates that may overcome the reliance of specific prompts (Shin et al., 2020; Liu et al., 2021), it has also been shown that PLMs may fail to understand simple features such as negation (Kassner and Schütze, 2020). Indeed, PLMs may even be biased towards the words on the prompt, and base their answers on this or other confounds (e.g. co-occurrences) instead of understanding the relation itself. For instance, in the previously-mentioned prompt 'The capital of Paris is —', capital is already mentioned in the prompt.

To address this, one possible solution is to rely on word analogies. In order to study the existence of a target relation between a pair of words (e.g. the 'country:capital' relation between Paris and France), one could simply put them together with another pair holding the target relation (e.g. Rome and Italy) in an analogy sentence (e.g. 'Paris to France is like Rome to Italy') and probe the model in a binary classification setting to check whether the analogy holds.[1]. Taking word analogies as the main reference, our aim is therefore to understand whether the transformer-based language models hold sufficient relational information to classify an analogy as being true or false. While from previous work we know that PLMs are indeed able to solve various types of analogies (Ushio et al., 2021b), we are interested in analyzing three different aspects for which we propose three probe tasks. In short, these probes are aimed at understanding the PLMs capability for (1) making fine-grained distinctions between concepts within the same types; (2) capturing the directionality of unidirectional relations; and (3) distinguishing types such as the difference

[1]This analogy probe is also grounded in research in cognitive psychology, where it has been argued that inferences made from analogies is a key feature to understand human creativity (Holyoak et al., 1996).

between capital and country.

2 Methodology

In this section, we describe our methodology to probe relational knowledge from language models through word analogies. Given two different tuples (h_1, t_1) and (h_2, t_2), and an analogy template T, we can generate an analogy sentence by inserting the head and tail words of tuples into their respective positions in T. An analogy sentence holds with respect to relation R if both tuples used in the generation process are members of the relation set R. For instance, the tuples $(Paris, France)$ and $(Rome, Italy)$ form a correct analogy for the 'country:capital' relation.

2.1 Probe Evaluation Setting

For our probes, we rely on two distinct settings which are usually linked to different usages of the language models, namely supervised and unsupervised (or *zero-shot*). The starting point for these two probe settings is a relation $R = \{(h_1, t_1), (h_2, t_2), (h_3, t_3), ...\}$ where (h_i, t_i) represents a tuple belonging to the given relation.

Supervised setting. Having the relation set R as input, we can frame the task as a binary classification where the input is a tuple and the output is *True* or *False* depending on whether the tuple belongs to the relation or not. For example, *Paris-France* would be a positive example for the *capital-of* relation, while *Paris-Rome* would be a negative example. In turn, R can be easily split between training and test sets, and negative samples can be obtained in different ways depending on the actual probe.

Unsupervised setting. In this setting, similarly obtained negative tuples can be paired with positive tuples. In this case, given an input pair (h_i, t_i) from R, and a pair of two additional tuples (a positive and a negative example), the task would consist of identifying the tuple better representing the relation R. For instance, given *Paris-France* and the tuples *Rome-Italy* and *Italy-France* as possible options, the correct answer would be *Rome-Italy*.

These two settings (i.e., supervised and unsupervised) provide additional insights in relation to two distinct theories with respect to relational knowledge. The supervised binary classification setting corresponds to the rigid theory in which relations either hold or not (Beckwith et al., 1991), present in resources such as WordNet (Miller, 1995). Stemming from cognitive psychology, the unsupervised comparative evaluation setting reflects on the graded assumption in which relations are a more fluid and it is not always possible to provide a clear binary distinction (Rosch, 1973). This comparative setting has been the basis to construct graded relational datasets (Vulić et al., 2017).

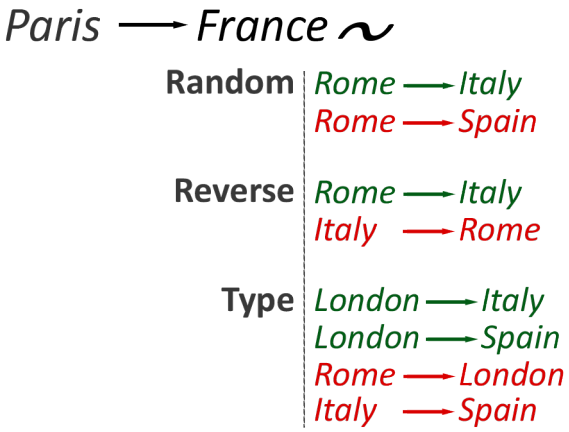

Figure 1: Sample positive and negative pairs for the three proposed probes given *Paris-France* as input.

2.2 Probes

We propose three probes to understand how language models capture three different aspects within relational knowledge. The probes mainly construct negative samples in a different way in order to test various features. The input in all cases is any given (h_i, t_i) in R. Figure 1 lists some sample positive and negative pairs for the three probes.

Random replacement. There are two ways to construct negative samples for this probe. A negative sample in this probe consists of (1) a tuple in R and an auxiliary tuple (h_j, t_i) in which $h_j \neq h_i$ (we refer to this probe as random-head); or (h_i, t_j) in which $t_i \neq t_j$ (random-tail). This probe aims at understanding how hard it is for the language model to identify a relation when the types of the head and tail are maintained.

Reverse direction. For the negative samples, the auxiliary tuple is simply a random tuple from R in which the positions of head and tail are reversed and follows the form (t_i, h_i). This probe aims at understanding to what extent PLMs understand the directionality of a given relation.

Type. For each input (h_i, t_i), as negative examples we take any two tuples (h_j, h_k) $(h_i \neq h_j \neq h_k)$ and (t_j, t_k) $(t_i \neq t_j \neq t_k)$. The goal of this probe is to test the capability of PLMs to understand the types of a given relation and, specifically, that different types are required for the relation to hold.

2.3 Datasets

We opted for the relation sets in Bigger Analogy Test Set (BATS) dataset (Gladkova et al., 2016). BATS has been shown to be more robust and complete to other analogy datasets such as Google-analogies (Mikolov et al., 2013). BATS covers a collection of 40 different inflectional and derivational morphology, lexicographic and encyclopedic relation sets. For each of these sets, we create six distinct datasets for each setting and probe.[2]

In the supervised setting, all datasets are initially split into training and test, with half of the tuples in each partition. Then, having an initial train/test partition R of n tuples, we can generate $n \times (n-1)$ positive by pairing each tuple with each other. In order to keep the dataset balanced, we also generate the same number of negative samples by following the probe methodologies described in the previous subsection. To form instances in the unsupervised dataset, we simply add a negative example to each positive instance from the supervised datasets.

2.4 Probe Architecture

For our supervised probe, we opted for RoBERTa-large and RoBERTa-base (Liu et al., 2019) as the PLMs in our experiments. In order to feed our dataset samples to these models, we make use of the analogy template "What — is to —, — is to — .", which was shown to be the most reliable general-purpose prompt for modelling analogies in Ushio et al. (2021b). We then pull the embedding of the target words and concatenate them together to form a larger feature vector which is ultimately fed into a simple multi-layer perceptron binary classifier.

For the unsupervised setting, we pick the option tuple that results in an analogy sentence with lowest pseudo-perplexity when inserted into our analogy template together with the query tuple. Given the tokenized form $[w_1, w_2, ..., w_{|S|}]$ of a sentence S, pseudo-perplexity is defined as:

$$\text{PPPL}(S) = \exp\left(-\sum_{i=1}^{|S|} \log P\left(w_i | S_{\setminus i}\right)\right) \quad (1)$$

in which $P\left(w_i | S_{\setminus i}\right)$ is the pseudo-likelihood (Wang and Cho, 2019) and $S_{\setminus i}$ is the tokenized form of S where the i-th token is replaced with a <mask> token.

3 Probe Evaluation

In this section, we present the results of our probe evaluation based on the methodology described in the previous section. First, we describe the embedding-based baselines in Section 3.1, and then we present the experimental results in Section 3.2.

3.1 Baselines

In order to put our results into perspective, we perform experiments using two embedding-based baselines using both relation and word embeddings. As relation embedding model we compare with RelBERT (Ushio et al., 2021a), a model specifically trained to extract relation embeddings from language models. Since RelBERT does not require the input tuples to be in a context, we can use the tuples without any analogy template. In the case of the unsupervised experiments, we extract the RelBERT relation embeddings of the input and candidate tuples, and choose the candidate tuple that has the embedding with highest cosine similarity to that of the input tuple. For the supervised setting, we simply feed the concatenation of the embedding vectors to a multi-layer perceptron binary classifier.

Similarly, we also report the results of a simple FastText-based (Bojanowski et al., 2016) static word embedding baseline. For this baseline, the relation embedding is obtained by simply computing the difference of individual word embeddings in a tuple, which is the standard pair encoding method used in the literature (Weeds et al., 2014; Vylomova et al., 2016; Camacho-Collados et al., 2019). Once this pair embedding is obtained, the rest of the methodology is the same as the one described for RelBERT.

3.2 Results

Table 1 shows our main experimental results. At first glance, semantic relations appear to be harder than morphological ones. When analysing model size, the larger RoBERTa model consistently outperforms its smaller counterpart, which goes in line

[2]Datasets are available in the supplementary material.

Type	Model	Setting					
		supervised			unsupervised		
		random	reverse	type	random	reverse	type
ES	RoB-Base	58.33	96.69	90.29	70.33	89.30	68.54
	RoB-Large	62.50	98.46	94.44	**74.92**	90.58	74.14
	RelBERT	**72.39**	98.68	**97.38**	63.55	**93.19**	**87.30**
	fastText	61.96	**99.71**	95.85	58.10	92.40	81.17
LS	RoB-Base	57.77	86.83	79.17	69.81	64.61	54.10
	RoB-Large	68.91	86.45	79.91	**78.43**	**68.77**	56.85
	RelBERT	**76.10**	82.80	**82.09**	70.97	68.60	**61.46**
	fastText	62.63	**90.20**	81.44	51.96	68.06	59.13
DM	RoB-Base	88.40	99.53	96.15	88.20	79.98	77.31
	RoB-Large	92.75	**99.65**	**96.34**	**97.43**	94.10	**84.39**
	RelBERT	**95.64**	97.14	95.73	91.69	89.22	70.20
	fastText	77.16	98.80	91.62	70.52	**95.98**	83.53
IM	RoB-Base	81.81	98.07	93.21	**96.83**	97.17	**87.76**
	RoB-Large	90.11	99.71	95.44	95.10	**98.44**	82.08
	RelBERT	**91.42**	98.33	**96.87**	92.78	92.30	68.05
	fastText	75.46	**99.80**	92.38	82.30	96.81	86.06
All	RoB-Base	71.58	95.28	89.70	81.29	82.76	71.93
	RoB-Large	78.57	96.07	91.53	**86.47**	87.97	74.36
	RelBERT	**83.89**	94.24	**93.02**	79.75	85.83	71.75
	fastText	69.30	**97.13**	90.32	65.72	**88.31**	**77.47**

Table 1: Average accuracy results of comparison models on our probe datasets grouped by general relation types (ES: Encyclopedic Semantics, DM: Derivational Morphology, LS: Lexicographic Semantics, IM: Inflectional Morphology). The last row includes the overall averaged results for all relations types.

with general language modelling results and in particular for modelling relations (Petroni et al., 2019). Regarding the supervised experiments, RoBERTa performs better in the reverse and type probes compared to random. This indicates the ability of PLMs (and in general distributional models given the strong fastText-based results) to capture word categories and their direction, while having room for improvement when it comes to capture more fine-grained distinctions proposed in the random probe.

In the unsupervised experiments the difference between relations is less marked in the case of PLMs. In this setting, except for the random probe, a simple word embedding baseline such as fastText prove more reliable. The superior performance on the reverse and type probes compared to random is more pronounced in the case of fastText baseline. This suggests that PLMs can capture more fine-grained meaning variances compared to static embeddings. Moreover, the comparable performance of fastText to the best performing models in unsupervised reverse and random probes indicates that contextual information encoded in contextualized embeddings, as opposed to the type/category information, play a less important role in these probing configurations.

4 Analysis

Since the goal of this paper is to probe PLMs for relational knowledge, for this extended analysis we focus on the supervised setting of the plain RoBERTa-large, which is more in line with most downstream applications.

Word Frequency First, we estimated the number of word occurrences in the underlying pre-training corpora[3], and following Chiang et al. (2020) we took the harmonic mean of occurrences of words in an output tuple as an estimate of tuple occurrence frequency. We hypothesized that most of the errors may be produced when the frequency of the output pair is low, as the RoBERTa may be less familiar with the words themselves. To this end, we computed a Kolmogorov-Smirnov for each relation type in which we separated the instances by correct and wrong decisions made by the model in the 'random' probe. For most relation types (over two thirds), p-values are higher than 0.05 for which we can conclude that frequency does not play a significant role in the performance. Those relation types where the effect seems more significant are *male-female* and *adj:comparative*.

Breakdown by Relation Table 2 shows a breakdown of the results by relation type. In general, we can observe poor performance on one/many-to-many type relations in 'random' and bidirectional relations in 'reverse'. This is an interesting sanity check which we would expect given the nature of the probes. In order to disentangle the effect that these relations may have in the final performance, we also computed the average accuracy excluding all one/many-to-many and bidirectional relation types. The main conclusions from Section 3.2 hold in which the 'random' probe appears to be harder than the others, with the average overall performance being 68.5 ('random'), 98.2 ('reverse') and 95.6 ('type'). Another interesting finding is the consistent superior performance on 'reverse' compared to 'type'. This is true in particular for the relations where head and tail words are coming from closer general categories (e.g. meronyms:part), which indicates that merely relying on word types may not be enough to capture directionality.

[3]Wikipedia+BooksCorpus (Zhu et al., 2015) was used as a proxy to get word frequencies. This was part of the full corpus where RoBERTa was trained on, which was not available.

	Relation	Semantic		
		random	reverse	type
Encyclopedic Semantics	male : female	73.86	95.90	90.02
	name : nationality	68.21	100	99.20
	country : language	67.25	100	98.68
	animal : sound	52.72*	98.94	90.88
	UK_city : county	59.50	100	95.42
	animal : shelter	51.72*	96.82	90.71
	things : color	52.91*	99.98	98.69
	animal : young	55.62*	96.78	89.80
	name : occupation	64.30	100	99.35
	country : capital	77.85	99.15	90.85
Lexicographic Semantics	meronyms : part	60.36*	92.35	78.06
	antonyms : binary	69.72*	56.90†	59.39
	synonyms : exact	69.30*	70.14†	66.53
	hyponyms : misc	63.43*	92.69	85.43
	antonyms : gradable	74.61*	90.00	81.41
	meronyms : member	67.66*	89.18	77.06
	hypernyms : misc	56.23*	99.35	96.49
	meronyms : substance	59.18*	87.86	77.14
	synonyms : intensity	71.12*	86.02	78.43
	hypernyms : animals	47.50*	99.88	97.69

	Relation	Morphological		
		random	reverse	type
Derivational Morphology	re+verb_reg	92.88	97.50	94.08
	verb+ment_irreg	95.85	100	94.92
	adj+ness_reg	96.12	100	97.83
	verb+tion_irreg	96.90	100	98.77
	verb+able_reg	92.19	100	97.31
	noun+less_reg	90.31	100	98.46
	adj+ly_reg	97.04	100	98.50
	un+adj_reg	97.54	100	98.67
	verb+er_irreg	90.79	99.83	94.00
	over+adj_reg	97.48	100	99.16
Inflectional Morphology	verb_inf : Ving	97.38	99.92	99.17
	verb_Ving : 3pSg	94.08	99.50	95.50
	verb_inf : Ved	96.67	100	93.75
	verb_inf : 3pSg	92.54	100	92.83
	noun : plural_reg	98.83	100	93.25
	noun : plural_irreg	96.63	98.00	89.92
	verb_3pSg : Ved	95.75	100	96.33
	verb_Ving : Ved	88.96	99.83	98.08
	adj : superlative	92.75	100	97.83
	adj : comparative	94.75	100	95.92

Table 2: RoBERTa-large results grouped by relation type (supervised setting). The results of one/many-to-many relations on the random probe and bidirectional relations on the reverse probe are marked by $*$ and †, respectively.

5 Conclusion

In this paper, we have presented three probes to understand to what extent PLMs (or any model in general) understand different aspects of the relations. In general, the 'random' probe proves the most challenging for PLMs, which is aimed at capturing some fine-grained information between the different types in a relation. In contrast, these models can accurately capture the aspects related to directionality and the word categories (or *types*) involved in a relation. When investigating the reasons of this discrepancy, we did not find a clear correlation between word frequency and performance for the 'random' probe, except for specific relation types. In general, however, based on our unsupervised experiments, PLMs seem to be better equipped to solve this probe when comparing between different pairs, even without task-specific training data.

6 Limitations

Our experiments are limited in various respects. First, the only language analysed is English, which limits the conclusions that can be taken with respect to other languages, especially those structurally different and from different families. Second, our experiments are based on a limited number of both models (which can additionally vary in size with potentially different conclusions) and configurations/prompts. While we follow standard practice, there are potentially configurations that have not been explored and could alter the significant of the results. Third, word analogies have been shown by previous research to be prone to external biases or confounding factors that can alter the results (Linzen, 2016; Gladkova et al., 2016; Nissim et al., 2020). We minimized this impact by proposing clear binary classification and comparative tasks, instead of the usual predictive framing in word analogies. Fourth, the data utilised corresponds to a single dataset, i.e. BATS. While this dataset was constructed so a wide variety of relations are covered, these are still limited in number (40) and biased towards certain categories. All in all, our study can be considered to be a first attempt to probe relational knowledge through word analogies, which appears to be a promising area for future work.

Acknowledgements

Jose Camacho-Collados is supported by a UKRI Future Leaders Fellowship.

References

Richard Beckwith, Christiane Fellbaum, Derek Gross, and George A Miller. 1991. Organized on psycholinguistic principles. *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, page 211.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *Proceedings of AAAI*.

Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2019. Relational word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3286–3296, Florence, Italy. Association for Computational Linguistics.

Hsiao-Yu Chiang, Jose Camacho-Collados, and Zachary Pardos. 2020. Understanding the source of semantic regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 119–131, Online. Association for Computational Linguistics.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the Student Research Workshop at NAACL*, pages 8–15.

Keith J Holyoak, Keith James Holyoak, and Paul Thagard. 1996. *Mental leaps: Analogy in creative thought*. MIT press.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too.

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*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of HLT-NAACL*, pages 746–751.

George A Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473.

Eleanor H Rosch. 1973. Natural categories. *Cognitive psychology*, 4(3):328–350.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Asahi Ushio, Jose Camacho-Collados, and Steven Schockaert. 2021a. Distilling relation embeddings from pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9044–9062, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021b. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. HyperLex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682, Berlin, Germany. Association for Computational Linguistics.

Jonas Wallat, Jaspreet Singh, and Avishek Anand. 2020. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third*

BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 174–183, Online. 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.

Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2249–2259, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.