Using Wikidata for Enhancing Compositionality in Pre-trained Language Models

Meriem Beloucif[1], **Mihir Bansal**[2], **Chris Biemann**[3]
[1]Uppsala University, [2]Carnegie Mellon University, [3]Universität Hamburg,
[1]meriem.beloucif@lingfil.uu.se, [2]mihirban@andrew.cmu.edu, [3]chris.biemann@uni-hamburg.de

Abstract

One of the many advantages of pre-trained language models (PLMs) such as BERT and RoBERTa is their flexibility and contextual nature. These features give PLMs strong capabilities for representing lexical semantics. However, PLMs seem incapable of capturing high-level semantics in terms of compositionality. We show that when augmented with the relevant semantic knowledge, PMLs learn to capture a higher degree of lexical compositionality. We annotate a large dataset from Wikidata highlighting a type of semantic inference that is easy for humans to understand but difficult for PLMs, like the correlation between age and date of birth. We use this resource for fine-tuning DistilBERT, BERT large and RoBERTa. Our results show that the performance of PLMs against the test data continuously improves when augmented with such a rich resource. Our results are corroborated by a consistent improvement over most GLUE benchmark natural language understanding tasks.

1 Introduction

Given their recent success in various natural language processing (NLP) tasks, there has been increasing work on understanding the abilities of pre-trained language models (PLMs) beyond what they can memorize. Having been trained on billions of words, BERT (Devlin et al., 2019) has shown impressive language representation abilities. However, there has not been much work on the degree of knowledge that BERT could infer about different topics from just the lexical information that they are trained on. Therefore, there has been a growing interest in probing PLMs on all kinds of linguistic, syntactic and semantic features (Huang et al., 2021; Beloucif and Biemann, 2021; Huang et al., 2021; Mosbach et al., 2020; Tenney et al., 2019; Peters et al., 2018b,a; Devlin et al., 2019; Radford and Narasimhan, 2018; Broscheit et al., 2022).

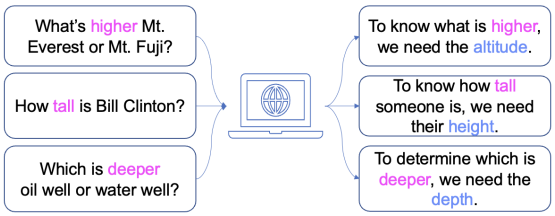

Figure 1: Multiple inferences are systematic for humans; however, they are much harder for NLP models to capture.

Figure 1 shows a few examples of high-level semantics relating to compositionality. For instance, when asked questions such as "What's higher Mt. Everest or Mt. Fuji?´´ or "How tall is Bill Clinton?´´, a person would most likely, and naturally think about *altitude* and *height* respectively, to accurately answer this question. When it comes to reasoning and inferences between semantic attributes (*net worth*) and their values (*rich*), humans can systematically infer between these concepts. The closer semantics in NLP that fits this case is compositional semantics since we investigate how different words in a sentence are linked to other words i.e. net worth being linked to wealth, and altitude is linked to the height of a mountain.

In this paper, we create a large dataset from Wikidata (Vrandečić and Krötzsch, 2014), where each sentence contains two words that are semantically related. We then fine-tune three pre-trained language models, namely BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019), using this data. We create test data that has the same style as the training data, but with different objects and inferences. We obtained a remarkable boost in the quality on the test data. Furthermore, we also report a consistent improvement over the GLUE benchmark for natural language understanding (Wang et al., 2018).

Our main contributions are:

Proceedings of Recent Advances in Natural Language Processing, pages 170–178
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_019

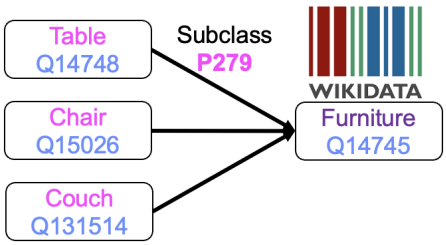

Figure 2: The hierarchical structure of Wikidata(Vrandečić and Krötzsch, 2014) allows us to have access to semantically sound data using different Wikidata entities as objects.

- a large dataset containing high-level semantics inferences,
- fine-tuning BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019) on a more semantically sensitive dataset using the masked language model predictions,
- improvements over the test data as well as the GLUE benchmark for natural language understanding.

2 Probing Pre-trained Language Models

Using probes has become a common way to investigate the knowledge encoded in transformer-based (Vaswani et al., 2017) pre-trained language models such as BERT. These investigations have varied from linguistic features to include commonsense knowledge and social biases that PLMs might have learned during the training. Wallace et al. (2019) used question answering to show that PLMs fail at rational reasoning when it comes to capturing the numerical commonsense. More work has focused on studying different linguistic features and the level of linguistic competence in different PLMs (Mosbach et al., 2020; Tenney et al., 2019; Peters et al., 2018b) by making use of fine-tuning and sentence-level semantics. Probes were also used to identify social toxicity and bias towards different interest groups as we as gender bias (Ousidhoum et al., 2021; Stanczak et al., 2021). Other probing experiments have been proposed to study the drawbacks of PLMs in areas such as the biomedical domain (Jin et al., 2019), syntax (Hewitt and Manning, 2019), semantic and syntactic sentence structures (Yenicelik et al., 2020; Tenney et al., 2019; Peters et al., 2018b), linguistics (Belinkov et al., 2017; Clark et al., 2020; Tenney et al., 2019) and commonsense knowledge (Petroni et al., 2019; Davison et al., 2019; Talmor et al., 2020). When it comes to language understanding, Yenicelik et al. (2020) showed that when it comes to polysemy, BERT creates closed semantic regions that are not clearly distinguishable from each other. Another finding relating to semantics (Beloucif and Biemann, 2021) conveys that, unlike syntax, semantics and general world knowledge are not inherently learned, and thus not brought to the surface by the representations obtained from pre-trained language models.

3 Data Creation

We use the knowledge graph extracted from Wikidata to construct the dataset. Wikidata (Vrandečić and Krötzsch, 2014) is a collaborative knowledge base, containing triples (entity id, property id, value id) that define a type of relation holding between an entity and a value. Wikidata also contains labels and aliases for the properties, entities, and values, which makes it the perfect resource for extracting similar objects that are likely to have similar values. We then investigate the ability of PLMs to capture the semantic relationship between the attribute-value pairs and further fine-tune PLMs to capture this relation effectively. [1]

Algorithm 1: Creating fine-tuning data from Wikidata objects.

Result: fine-tuning dataset
fine-tuning-data=; **while** *keyword in (food, furniture, city, tool)* **do**
  AllData=extract all subclasses of keyword from Wikidata,
end
while *i=0, i<size(alldata), i++* **do**
  $BERT - sent(i)$= BERT prediction on sentence i, extract all subclasses of keyword from Wikidata,
  if *BERT-sent(prediction) == accurate-prediction* **then**
    fine-tuning-data=fine-tuning-data + BERT-sent(i)
  else
end

In the knowledge graph, we focused on entities that were labeled *food*, *furniture*, *city* and *tool*, with *nutritious-healthy*, *wider-width*, *rainfall-humidity* and *longer-length* as entity-value pair respectively.

[1]The final dataset and the code are available here: https://github.com/mihir86/Fine-Tuning-BERT-with-Wikidata

Model	Top Prediction Accuracy				Top 5 Prediction Accuracy			
	PTLM	one-word-fine-tuned	two-word-fine-tuned	all-words-fine-tuned	PTLM	one-word-fine-tuned	two-word-fine-tuned	all-words-fine-tuned
DistilBERT-base	24%	62%	56%	**66%**	46%	**96%**	94%	92%
RoBERTa-large	20%	**38%**	**38%**	8%	44%	**78%**	70%	42%
BERT-Large	0%	**26%**	24%	22%	0%	**42%**	36%	40%

Table 1: The Performance of BERT on the test data.

Food is selected as the key because food items exhibit the attribute of *nutrition*, and thus comparing the subclasses of *food*, in terms of their nutrition can enable us to compare which food item is more *healthy*. For *city*, different cities have different *rainfall* and thus comparing the *rainfall* between different subclasses and instances of *city* can enable us to compare which city has more *humidity*. We applied the same analysis to *furniture* and *tool*.

In order to capture the semantic relationship between the attribute-value pairs, we create a dataset from the sentences where the value in the attribute-value relationship had been accurately predicted by BERT. The subclasses and instances of the keys *food*, *furniture*, *city* and *tool* were extracted from the knowledge graph and then used in combination with each other to create sentences of the form "Which is [attribute], and thus has more [value], [object 1] or [object 2]´´ where the objects represent the words used for comparing the attribute-value pair. For example, to analyze the ability of PLMs to capture the semantic relationship between *wider(attribute)* and *width(value)*, we consider *bed(Object 1)* and *chair(Object 2)* to be the chosen subclass combinations of the key *furniture*. Therefore, the sentence "Which is wider, and thus has more width, bed or chair?´´, is constructed with *width(value)* being masked.

Our final dataset contains around 8,000 fine-tuning samples, using five distinct attribute-value pairs. We divided our data into three categories, a dataset containing: (1) one-word objects, such as chairs, and couscous; (2) one-word objects and two-word compounds, such as folding chairs and bean sprout; and (3) all possibilities, including three-word compounds, such as aged cheddar cheese and slip joint plier. The purpose is to check how compound words affect the accuracy of the fine-tuned model, or in other words, does it matter to the PLM whether a noun is a compound or not?

4 Fine-Tuning PLMs for High-level Semantics

We used Huggingface(Wolf et al., 2019, 2020) for fine-tuning BERT (Devlin et al., 2019), RoBERTa(Liu et al., 2019) and DistilBERT(Sanh et al., 2019) [2]. For the fine-tuning, 15% of the tokens were masked randomly and the PLMs are fine-tuned with a masked language model objective by minimizing the loss based on the gold standard. The fine-tuned model is then evaluated on the test dataset, which consists of 50 different sentences with different semantic relationships.

5 Experimental Setup

Test data When finetuning the PLMs, one of the most challenging tasks is to prove that model could learn from the finetuning and is not just overfitting to the specific task. For that reason, we are testing on two different datasets: A Wikidata-based test set and the GLUE benchmark for natural language understanding (Wang et al., 2018). BERT-based models have significantly increased state-of-the-art over the GLUE benchmark, and most of the best scoring models for this benchmark include or elaborate on BERT.

We train our model on five topics, with different objects, but we test on 50 different attribute-value pairs. In order to show a certain generalization over the training data, we made sure that no attribute-value pair from the training is part of the test data. The masked word is then predicted by different PLMs. The accuracy of the top one and top five predictions is calculated. We purposefully diversify our test set from our training set to show that the improvement is not mere memorization. Our test data contains different objects such the *Eiffel Tower* or *Burj Khalifa*, which are both instances

[2] https://huggingface.co/models

Model	Score	CoLA	MNLI (M/MM)	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
DistilBERT	75.3	47.2	80.8 / 82.0	85.6	88.2	85.6	52.7	90.4	84.1	56.3
DistilBERT-FT	**76.0**	**49.9**	80.8 / 81.8	**87.1**	**88.4**	85.5	**56.3**	90.1	**85.0**	54.9
RoBERTa	83.5	63.6	90.2 / 90.2	91.4	93.8	92.2	71.2	95.3	91.7	55.3
RoBERTa-FT	**83.6**	**64.7**	89.4 / 89.2	**91.5**	**94.1**	**92.4**	**72.6**	95.0	**92.6**	54.9
BERT	79.5	60.5	86.7 / 85.9	89.3	92.7	72.1	70.1	94.9	86.5	56.3
BERT-FT	**80.3**	**61.1**	86.6 / **86.5**	**90.9**	**93.6**	**72.4**	**72.9**	92.4	**90.2**	56.3

Table 2: The Performance of all three models on the GLUE benchmark.

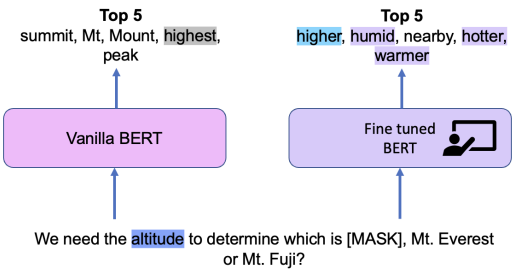

Figure 3: BERT cannot predict the correct predictions when it comes to the mountain context. After fine-tuning, the predictions are more relevant to the context, even though *altitude* was not part of the fine-tuning data.

of the subclasses *observation tower* and *tourist attractions*. We report the accuracy (Table 1) in two distinct cases: (1) is the top one prediction correct?; and (2) is the correct prediction within the top five predictions?

Results Table 1 shows the prediction accuracy for all three models, before and after fine-tuning. The performance gain is consistent across the top one prediction and the top five predictions. We note from Table 1 that DistilBERT has the highest improvement compared to RoBERTa and BERT large. RoBERTa and BERT large are more sensitive to compound words, and they perform best with the one-word object and two words object. For the top five prediction accuracy, all three models perform best without compound words. In Figure 3 we show a concrete example from DistilBERT fine-tuning. We note from the example that, even though *altitude* is not part of the fine-tuning dataset, PLMs are now able to generalize from the concepts, rather than just memorize the words.

Testing on the GLUE benchmark corroborates this finding even further. Table 3 shows a significant improvement for some tasks and a slight improvement on other tasks. More specifically, we note that for the single task datasets, such as the Corpus of Linguistic Acceptability, CoLa (Warstadt et al., 2019), and for The Stanford Sentiment Treebank, SST-2 (Socher et al., 2013) there is a significant gain for the fine-tuned models. The same applies to inference tasks; Microsoft Research Paraphrase Corpus, MRPC, the Quora Question Pairs datasets, and the Semantic Textual Similarity Benchmark, STS-B (Cer et al., 2017), achieve a similar improvement. The consistent improvement over the semantically driven tasks shows that our fine-tuning helps PLMs capture more high-level semantics.

6 Conclusion

In this paper, we investigate how PLMs capture a very specific type of compositionality between different concepts. We also finetune two different PLMs on five different attribute-value pairs and test the model on 50 annotated themes. The training data and the test data have different topics and wording. Additionally, we purposefully limited the fine-tuning data for the scope of this short paper, since we did not want to make PLMs memorize all possible concepts. Our results show that, by having a resource that contains a basic level of lexical compositionality, we indeed help improve PLMs accuracy. However, we also show that there is more improvement in the GLUE tasks that are more semantically sensitive.

7 Limitations

Compositionality is a strong human characteristic when it comes to languages. In this paper, we created a synthetic dataset in order to help PLMs learn high-level semantics compositionality. The main limitation is the difficulty to test all possible cases. Compositionality is a challenging task, we show that we are able to generalize over limited test

data, however, given their complex architecture, it is challenging to make test generalizations in the human sense. The second point is related to the created dataset, although widely accepted in the field, synthetic data suffers from human authenticity. More specifically, in an everyday conversation, when a person is asked about their age, the deduction in the human brain is automatic. It is challenging to present that concept through a sentence, which is what we tried to do here for testing and enabling the finetune.

References

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 861–872, Vancouver, Canada.

Meriem Beloucif and Chris Biemann. 2021. Probing pre-trained language models for semantic attributes and their values. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2554–2559, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Broscheit, Quynh Do, and Judith Gaspers. 2022. Distributionally robust finetuning BERT for covariate drift in spoken language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1985, Dublin, Ireland. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1173–1178, Hong Kong, China.

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*, pages 4171–4186, Minneapolis, MN, USA.

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*, pages 4129–4138, Minneapolis, MN, USA.

James Y. Huang, Kuan-Hao Huang, and Kai-Wei Chang. 2021. Disentangling semantics and syntax in sentence embeddings with pre-trained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1372–1379, Online. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89, Minneapolis, MN, USA.

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. *CoRR*.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 68–82, Online. Association for Computational Linguistics.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. 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)*, Online. Association for Computational Linguistics.

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

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium.

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*, pages 2463–2473, Hong Kong, China.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Karolina Stanczak, Sagnik Ray Choudhury, Tiago Pimentel, Ryan Cotterell, and Isabelle Augenstein. 2021. Quantifying gender bias towards politicians in cross-lingual language models. *CoRR*, abs/2104.07505.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, pages 743–758.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the Association for Computing Machinery*, pages 78–85.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. 2020. How does BERT capture semantics? a closer look at polysemous words. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 156–162, Online. Association for Computational Linguistics.

A Appendix A

We use the AdamW optimizer along with a learning rate of 1e-4 and a batch size of 16 for fine-tuning. We perform the fine-tuning experiment with 2,3 and 4 epochs and with different varieties of datasets ranging from 'one-word', 'two-word' and 'all-words' cuts inside the dataset created.

Sentence	PTLM Predictions	Fine tuned Predictions
We need the altitude to determine which is [MASK], Mt. Everest or Mt. Fuji.	summit, Mt, Mount, highest, peak	**higher**, humid, nearby, hotter, warmer
Which is taller and thus has more [MASK], Eiffel Tower or Burj Khalifa?	seats, windows, rooms, room, wings	rainfall, width, **height**, weight, mass
We need the height to determine who is [MASK], Dwight D. Eisenhower or Bill Clinton	tallest, **taller**, tall, seated, correct	**taller**, tall, seated, correct, next
Rock is heavier, thus has a higher [MASK].	density, **weight**, yield, hardness, content	**weight**, rainfall, density, mass, temperature
This road is wider, thus it has more [MASK].	lanes, traffic, curves, bends, access	**width**, length, traffic, rainfall, weight
Which is deeper, and thus has more [MASK], swimming pool or ocean?	water, pool, pools, **depth**, amenities	**depth**, rainfall, width, length, depths
Which is deeper, and thus has more [MASK], oil well or water well?	wells, water, **depth**, well, reservoirs	**depth**, rainfall, width, depths, deeper
Who was born earlier, and is thus [MASK], Narendra Modi or Rahul Gandhi?	named, called, unknown, mentioned, identified	**older**, younger, more, healthy, born

Table 3: Examples of Semantic Improvement through fine-tuning. The examples are extracted from the test set.

Model	Iterations	Training loss	Top Prediction Accuracy	Top 5 Prediction Accuracy
DistilBERT-base-cased	2	0.015	58%	96%
RoBERTa-large	2	0.107	24%	70%
BERT-Large	2	0.397	22%	36%
DistilBERT-base-cased	3	0.0237	62%	96%
RoBERTa-large	3	0.114	38%	78%
BERT-Large	3	0.171	28%	36%
DistilBERT-base-cased	4	0.01	56%	90%
RoBERTa-large	4	0.124	34%	64%
BERT-Large	4	0.171	26%	42%

Table 4: Performance of BERT Fine-tuned with single word combinations in Wikidata on Test Dataset.

Model	Iterations	Training loss	Top Prediction Accuracy	Top 5 Prediction Accuracy
DistilBERT-base-cased	2	0.0644	56%	94%
RoBERTa-large	2	0.00407	38%	70%
BERT-Large	2	0.365	24%	36%
DistilBERT-base-cased	3	0.012	44%	86%
RoBERTa-large	3	0.165	30%	64%
BERT-Large	3	0.17	24%	36%
DistilBERT-base-cased	4	0.0382	46%	84%
RoBERTa-large	4	0.0447	32%	54%
BERT-Large	4	3.35	0%	0%

Table 5: Performance of BERT Fine-tuned with single and two word combinations in Wikidata on Test Dataset.

Model	Iterations	Training loss	Top Prediction Accuracy	Top 5 Prediction Accuracy
DistilBERT-base-cased	2	0.0628	66%	92%
RoBERTa-large	2	0.154	8%	42%
BERT-Large	2	0.394	22%	40%
DistilBERT-base-cased	3	0.0261	70%	88%
RoBERTa-large	3	3.24	0%	0%
BERT-Large	3	0.13	18%	38%
DistilBERT-base-cased	4	0.0604	68%	86%
RoBERTa-large	4	3.12	0%	0%
BERT-Large	4	0.0404	18%	40%

Table 6: Performance of BERT Fine-tuned with all combinations in Wikidata on Test Dataset.