Can Pre-trained Language Models Interpret Similes as Smart as Human?

Qianyu He[1,*], **Sijie Cheng**[1,*], **Zhixu Li**[1,†], **Rui Xie**[3], **Yanghua Xiao**[1,2,†]
[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[2]Fudan-Aishu Cognitive Intelligence Joint Research Center, Shanghai, China
[3]Meituan, Shanghai, China
qyhe21@m.fudan.edu.cn, rui.xie@meituan.com,
{sjcheng20, zhixuli, shawyh}@fudan.edu.cn

Abstract

Simile interpretation is a crucial task in natural language processing. Nowadays, pre-trained language models (PLMs) have achieved state-of-the-art performance on many tasks. However, it remains under-explored whether PLMs can interpret similes or not. In this paper, we investigate the ability of PLMs in simile interpretation by designing a novel task named Simile Property Probing, i.e., to let the PLMs infer the shared properties of similes. We construct our simile property probing datasets from both general textual corpora and human-designed questions, containing 1,633 examples covering seven main categories. Our empirical study based on the constructed datasets shows that PLMs can infer similes' shared properties while still underperforming humans. To bridge the gap with human performance, we additionally design a knowledge-enhanced training objective by incorporating the simile knowledge into PLMs via knowledge embedding methods. Our method results in a gain of 8.58% in the probing task and 1.37% in the downstream task of sentiment classification. The datasets and code are publicly available at https://github.com/Abbey4799/PLMs-Interpret-Simile.

1 Introduction

A simile is a figure of speech comparing two fundamentally different entities via shared properties (Paul, 1970). There are two types of similes as illustrated in Figure 1, *closed* similes explicitly reveal the shared properties between the topic entity and the vehicle entity, such as the property *"slow"* shared by *"lady"* and *"snail"* in the sentence *"The old lady walks as slow as a snail"*; while *open* similes do not state the shared property such as the sentence *"The old lady walks like a snail"*. Similes play a vital role in human expression to make literal utterances more vivid and graspable and are widely used in the corpus of various domains (Liu et al., 2018; Chakrabarty et al., 2020a; Zhang et al., 2020). It is estimated that over 30% of the comparisons can be regarded as similes in product reviews (Niculae and Danescu-Niculescu-Mizil, 2014).

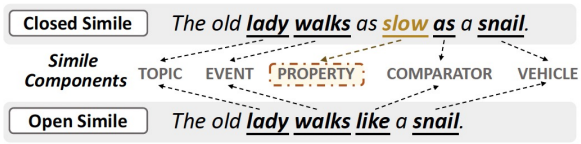

Figure 1: Examples of two types of similes. Whether the component *property* is stated determines the type of simile.

Simile interpretation is a crucial task in natural language processing (Veale and Hao, 2007; Qadir et al., 2016; Chakrabarty et al., 2021a), which can assist several downstream tasks such as understanding more sophisticated figurative language (Veale and Hao, 2007) and sentiment analysis (Niculae and Danescu-Niculescu-Mizil, 2014; Qadir et al., 2015). Take the simile *"the lawyer is like a shark"* for an example. Although all words in this simile are neutral, this simile expresses a negative affect since *"lawyer"* and *"shark"* share the negative property *"aggressive"*.

In the past few years, large pre-trained language models (PLMs) have achieved state-of-the-art performance on many natural language processing tasks (Devlin et al., 2018; Liu et al., 2019b). Recent studies suggest that PLMs have possessed various kinds of knowledge into contextual representations (Goldberg, 2019; Petroni et al., 2019; Lin et al., 2019; Cui et al., 2021). However, the ability of PLMs to interpret similes remains under-explored. Although some recent work (Chakrabarty et al., 2021a) studies the ability of PLMs in choosing or generating the plausible continuations in narratives, this way cannot fully reveal the ability of PLMs to interpret similes.

In this paper, we propose to investigate the abil-

*Equal contribution
†Corresponding author

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7875 - 7887

Category	Question Example	%
Qualities	My *client* is as `[MASK]` as a newborn *lamb* . **A.** innocent **B.** delicious **C.** legal **D.** guilty	27.78
Condition	The *toddler* was running around as `[MASK]` as a *bee*. **A.** busy **B.** yellow **C.** idle **D.** messy	22.28
Sense	His *anger* was as `[MASK]` as a burning *ember*. **A.** hot **B.** red **C.** cold **D.** warm	17.20
Measurement	My new baby *brother* is as `[MASK]` as a *button*. **A.** red **B.** tiny **C.** cute **D.** hot	14.16
Color	He was scared so much. *He* was as `[MASK]` as a *ghost*. **A.** white **B.** holy **C.** gay **D.** black	06.75
Time	The old *man* walks as `[MASK]` as a *tortoise*. **A.** young **B.** little **C.** slow **D.** quick	06.57
Emotion	The *boy* was as `[MASK]` as a *dog* that lost its bone. **A.** happy **B.** friendly **C.** sad **D.** glad	05.26

Table 1: Percentage and examples for our simile probes of different categories. The option marked with "___" indicates the correct answer. The italicized words one by one in each sentence are the topic, masked property, and vehicle, respectively.

ity of PLMs in simile interpretation by designing a novel task named as *Simile Property Probing*, i.e., to let the PLMs infer the shared properties of similes. Specifically, we design a particular masked-word-prediction probing task in the form of multiple-choice questions. This probe masks the explicit property of a *closed* simile and then lets the PLMs discriminate it from three distractors. To make the questions convincing and challenging, the distractors should be not only *true-negative* as they would introduce logical errors once they are filled in the sentence, but also *challenging* as they are semantically close to the correct answer. To achieve this, we propose to obtain some similar properties of the golden one from ConceptNet (Liu and Singh, 2004) and COMET (Bosselut et al., 2019), from which we select the three best distractors according to their proximity to the golden property in the feature space. From two different types of data sources: textual corpus collection and human-designed questions, we collect a total of 1,633 probes with various usage frequencies and context diversities, covering seven categories as listed in Table 1.

Based on our designed task, we evaluate the ability of BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019b) to infer the shared properties of similes. We perform an empirical evaluation in two settings: (1) zero-shot, where the models are off-the-shelf; (2) fine-tuned, where the models are fine-tuned with MLM objective via masking properties. We observe that PLMs have been able to infer properties of similes in the pre-training stage and the ability can be further enhanced by fine-tuning. However, fine-tuned PLMs still perform worse than humans. Moreover, we find that the simile components *vehicle* and *topic* contribute the most when inferring the properties.

Inspired by the sufficient hints offered by the components *vehicle* and *topic* in our empirical study, we propose a knowledge-enhanced training objective to further bridge the gap with human performance. Considering *property* (p) as the **relation** between *topic* (t) and *vehicle* (v), we design a simile knowledge embedding objective function following conventional knowledge embedding methods (Bordes et al., 2013) to incorporate the simile knowledge *(t,p,v)* into PLMs. To integrate simile knowledge and language understanding into PLMs, we jointly optimize the knowledge embedding objective and the MLM objective in our design. Overall, the knowledge-enhanced objective shows effectiveness in our probing task and the downstream task of sentiment classification.

To summarize, our contributions are three-fold: (1) To our best knowledge, we are the first to systematically evaluate the ability of PLMs in interpreting similes via a proposed novel simile property probing task. (2) We construct simile property probing datasets from both general textual corpora and human-designed questions, and the probing datasets contain 1,633 examples covering seven main categories of similes. (3) We also propose a novel knowledge-enhanced training objective by complementing the MLM objective with the knowledge embedding objective. This method gains 8.58% in the probing task and 1.37% in the downstream task of sentiment classification.

2 Preliminaries on Simile

A sentence of simile generally consists of five major components (Hanks, 2013; Niculae and Danescu-Niculescu-Mizil, 2014), where four are necessary and the remaining one is optional. The four explicit components are as follows: (1) **topic (or tenor)**: the subject of the comparison acting as

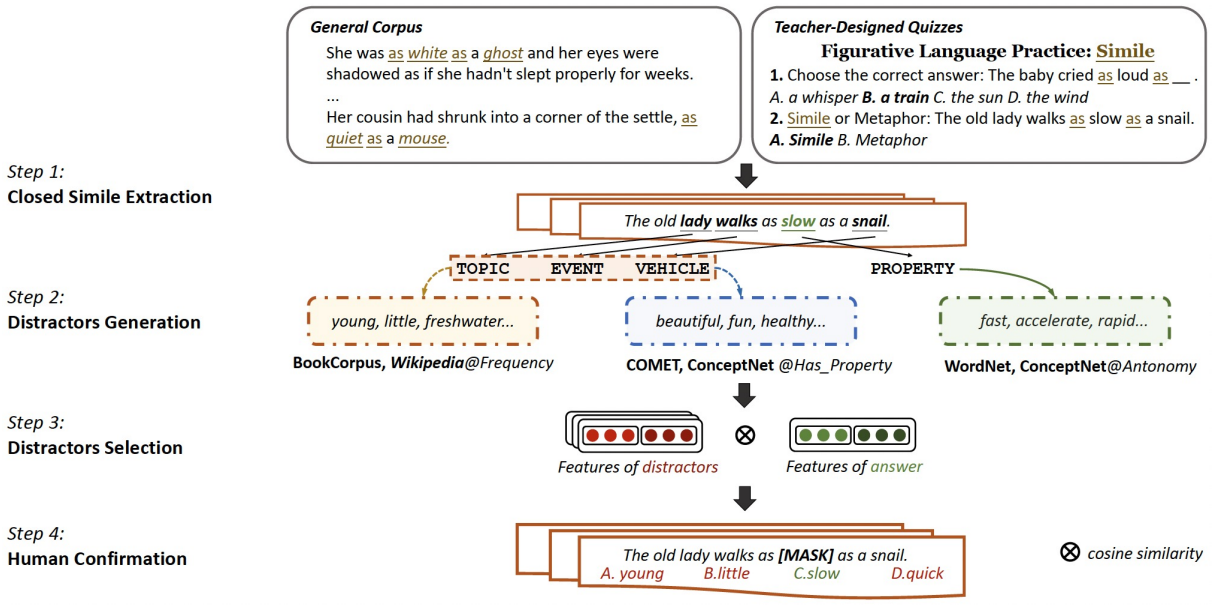

Figure 2: A process for designing our simile property probing task. In Step 1, we collect closed similes from two different sources. In Step 2, according to four important components in each simile, we generate distractor candidates with three strategies. In Step 3, we adopt cosine similarity to select more challenging distractors. In Step 4, we ask human annotators to ensure the quality and obtain our final probing datasets.

source domain; (2) **vehicle**: the object of the comparison acting as target domain; (3) **event**: the predicate indicating act or state; (4) **comparator**: the trigger word of a simile such as *as* or *like*. The optional component **property** reveals the shared characteristics between the topic and the vehicle. There are two types of similes depending on whether the property is explicit or implicit (Beardsley, 1981). The similes which mention the property directly are named as the *closed* similes, while the others are *open* similes, as shown in Figure 1.

3 The Simile Property Probing Task

3.1 Task Formulation

To estimate the ability of PLMs in simile interpretation, we design a particular *Simile Property Probing* task, which masks the explicit property of a *closed* simile, and then lets the PLMs discriminate it among four candidates. Considering that the shared properties between topic and vehicle may not be unique (Lacroix et al., 2005), we specifically design a multiple-choice question answering task (with only one correct answer) rather than a cloze task to probe the ability of PLMs to infer properties of similes, since the latter one may result in multiple correct answers.

Formally, given a simile text sequence $S = (w_1, w_2, ..., w_{i-1}, \texttt{[MASK]}, w_{i+1}, ..., w_N)$, where the shared property w_i between the topic and vehicle is masked, the probing task requires the PLMs to find the correct property from four options, where the other three options are hard distractors.

3.2 Probing Data Collection

We construct datasets for the proposed probing task in four steps. The overview of our probing data collection process is described in Figure 2.

3.2.1 Data Sources

We construct our datasets from two different sources to detect the capability of PLMs from two perspectives: textual corpus collection and human-designed questions. To avoid laborious human labeling on the implicit properties of open similes, we collect closed similes with explicit properties.

General Corpus. Following (Hanks, 2005; Niculae and Yaneva, 2013), we adopt two general corpora, British National Corpus (BNC)[1] and iWeb[2]. To identify closed similes, we extract the sentences matching the syntax *as ADJ as (a, an, the) NOUN*. Through syntactic pattern matching, we finally collect 1,917 sentences.

Teacher-Designed Quizzes. Questions about similes designed by teachers from educational resources are ideal sources for assessing the ability to understand similes. Hence, we choose Quizizz[3], an

[1]https://www.english-corpora.org/bnc/
[2]https://www.english-corpora.org/iweb/
[3]https://quizizz.com/

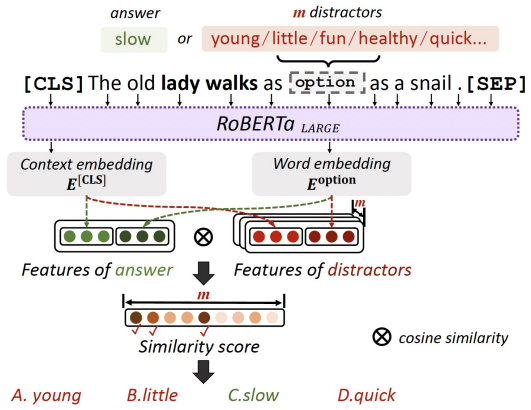

Figure 3: Illustration of the distractor selection method.

Dataset	General Corpus	Quizzes
#Sentence	775	858
#Unique *topic* concept	415	366
#Unique *property* concept	280	160
#Unique *vehicle* concept	522	250
#Unique *event* concept	147	66
#Unique *topic-vehicle* pair	743	684
#Unique *topic-property-vehicle* pair	751	701
Maximum sentence length	98	44
Average sentence length	25.80	12.69
Minimum sentence length	7	7
@Start	34.32%	20.40%
@Middle	43.23%	63.29%
@End	22.45%	16.32%

Table 2: Statistics of our simile property probing datasets. @ denotes the position of the simile in the given sentence.

emerging learning platform founded in 2015. On this platform, users can create quizzes on a specific topic as teachers to assess students' understanding of related knowledge. We collect a set of quizzes with titles concerning similes and extract the complete closed simile sentences from the questions and answers in these quizzes. Finally, we retrieve 875 complete closed similes from 1,235 quizzes.

To assure the quality of our constructed datasets and prepare for further analysis, three annotators are required to decide whether the extracted sentences are similes or not, and annotate their corresponding simile components. The inter-annotator agreement on identifying similes is 0.77 using Fleiss' Kappa score (Fleiss, 1971). All the properties in our datasets are single-token by replacing multi-token properties with their single-token synonyms in the knowledge base WordNet (Miller, 1995) and ConceptNet (Liu and Singh, 2004).

3.2.2 Distractor Design

To make our probes convincing, three distractors are designed against the original property in each simile with two criteria (Haladyna et al., 2002; Ren and Zhu, 2020): *true-negative* and *challenging*. We argue that well-designed distractors should be illogical when filled into the questions *(true-negative)* while being semantically related to the correct answer *(challenging)*. Our distractor design mainly involves three phases: 1) distractor generation; 2) distractor selection; 3) Human Confirmation.

Distractor Generation. To meet the requirement of *challenging*, we generate distractor candidates from the four semantic-related components of a simile, i.e., topic, vehicle, event, and property. Given the original property, we harvest its antonyms from the knowledge base WordNet and ConceptNet. With regard to three other components, we extract their properties from two sources as follows. Given a component, we utilize the *HasProperty* relation from ConceptNet (Liu and Singh, 2004) and COMET (Bosselut et al., 2019) to retrieve the property. Moreover, we rank the adjectives or adverbs concerning[4] each component in Wikipedia and BookCorpus corpus[5] by frequency and select the top ten candidates with a frequency of more than one.

Distractor Selection. To select the most *challenging* distractors from the generated distractor candidates, we propose to measure the similarity between the original sentence with the correct property and the sentence with a distractor. Intuitively, the more similar the two sentences, the more challenging the distractor. An example of the distractor selection process is depicted in Figure 3. Given the original sentence or the new sentence replacing the correct property with a distractor, we first utilize RoBERTa$_{\text{LARGE}}$ to extract two types of features. One feature is context embedding, which is the sentence embedding of `[CLS]`, while the other feature is word embedding, which is the token embedding of the answer or distractors. We then concatenate the embeddings of the two features to compute the cosine similarity between sentences with the answer and a distractor. Finally, we select the top 3 distractors with the highest similarities.

Human Confirmation. To ensure the distractors are *true-negative*, three human annotators are asked to label each selected distractor. If more than two annotators are uncertain about its correctness, we replace it with another suitable candidate.

[4]We adopt dependency parsing via the StanfordNLP tool to find adjectives and adverbs related to components.

[5]https://huggingface.co/datasets/

3.2.3 Statistics of the Datasets

Table 2 presents the statistics of our constructed datasets. We count unique components and component pairs to present the usage frequencies of similes. The length of the sentences in each dataset indicates the diversities of context. Additionally, we analyze the distribution of the position of simile in the sentences in each dataset, where *start*, *middle* and *end* correspond to the positions of the three equally divided parts of each sentence. We also investigate the categories covered by our datasets. The results and details about the category classification are provided in Appendix C. Overall, the Quizzes dataset provides similes commonly expressed by people, while the General Corpus dataset presents similes with more diverse contexts.

3.3 Supervision for Fine-Tuning PLMs

Besides evaluating the ability of PLMs in the zero-shot setting where the models are off-the-shelf, we also study whether the performance could be improved through fine-tuning with the MLM objective via masking properties. To achieve this, we collect training data from Standardized Project Gutenberg Corpus[6] (SPGC) (Gerlach and Font-Clos, 2020). SPGC is a 3 billion words corpus collected from about 60 thousand eBooks. We extract similes via matching the syntactic pattern *(Noun ... as ADJ as ... NOUN)* and end up with 4,510 sentences. Additionally, we adopt dependency parsing[7] to automatically annotate the simile components of each sentence without human labor.

4 Empirical Study on PLMs

In this section, we first conduct a set of experiments to probe the ability of PLMs to infer properties in similes and then evaluate the influence of each component on the model performance.

4.1 Ability to Infer Shared Properties

4.1.1 Experiment Set-up

To disentangle what is captured by the original representations and what is introduced from fine-tuning stage, we apply two different types of settings: (1) zero-shot; (2) fine-tuning. In our first setting, we use BERT and RoBERTa with pre-trained masked-word-prediction heads to perform our probing task. In the second setting, we utilize the MLM training objective inherited from PLMs to fine-tune the models. We replace the property of each simile with the special token `[MASK]` in our constructed supervised datasets (Section 3.3) and ask models to recover the original property. The experimental details are provided in the Appendix B.

We mainly compare the model accuracy of PLMs with the following baselines: (1) **EMB** (Qadir et al., 2016): It obtains the composite simile vector by performing an element-wise sum of the word embedding for the vehicle and event, then selects the answer with the shortest cosine distance from the composite vector. (2) **Meta4meaning** (Xiao et al., 2016) : This method prefers the properties which are strongly associated with both topic and vehicle. It also prefers the properties that are more relevant to the vehicle than to the topic. The association is measured by statistical significance. (3) **ConScore** (Zheng et al., 2019) : It suggests that better properties would have a smaller and balanced distance to the topic and vehicle in the word embedding space. (4) **MIUWE** (Bar et al., 2020) : The ranking method assigns each property a list of scores, including the statistical co-occurrences and similarity to the collocations of the topic and vehicle. The baselines above mainly consider the statistical information and embedding similarities between the properties and the simile components. The other baseline is human performance. We sample 250 random questions from both datasets, and for each question, we gather answers from three people. We take the majority vote as the human performance of our probing task and ensure that three annotators agree on the questions that they gave completely different annotation results.

Setting	Models	General Corpus	Quizzes	Gain
	ConScore (Zheng et al., 2019)	27.48	34.85	-
	Meta4meaning (Xiao et al., 2016)	27.74	47.44	-
	EMB (Qadir et al., 2016)	28.27	47.90	-
	MIUWE (Bar et al., 2020)	30.97	53.85	-
Zero-Shot	**BERT$_{BASE}$**	64.13	74.36	-
	BERT$_{LARGE}$	72.39	83.22	-
	RoBERTa$_{BASE}$	69.55	82.87	-
	RoBERTa$_{LARGE}$	78.97	87.41	-
Fine-tuned	**MLM-BERT$_{BASE}$**	67.74	82.05	+5.65
	MLM-BERT$_{LARGE}$	73.85	84.58	+1.40
	MLM-RoBERTa$_{BASE}$	70.58	84.69	+1.43
	MLM-RoBERTa$_{LARGE}$	**78.97**	**88.97**	+0.78
	Human Performance	87.60	93.60	-

Table 3: Accuracy of different models in our simile property probing task.

4.1.2 Results

The accuracies of different methods under two different settings on our datasets are listed in Table 3,

[6] https://github.com/pgcorpus/gutenberg/
[7] https://stanfordnlp.github.io/CoreNLP/

where the last column represents the average absolute gains of each PLM after fine-tuning with the MLM objective. All the results of our experiments are averaged over three random seeds. First of all, the prediction accuracies of both BERT and RoBERTa in the zero-shot setting are much higher than the baselines only considering the statistical information and embedding similarities between simile components. This phenomenon indicates that the knowledge learning from the pre-train stage can help infer the simile properties. Moreover, the performance can be further improved by training with the MLM objective, demonstrating that the fine-tuning phase with the supervised dataset can introduce related knowledge about similes. However, models still underperform humans by several accuracy points, leaving room for improvement in our probing task.

Overall, all the models perform better on Quizzes Dataset than on General Corpus Dataset, indicating that more diverse contexts increase the difficulty of inferring the shared properties. Also, RoBERTa consistently outperforms BERT, likely due to a larger pre-training corpus containing more similes. More complementary results are provided in the Appendix A.1.

4.2 Influence of Important Components

4.2.1 Experiment Set-up

Due to the high performance of off-the-shelf PLMs, we are interested in the contributions of each component to infer shared properties in the zero-shot setting. First, the information of each component is hidden through a certain strategy. Specifically, for *topic*, *vehicle* and *comparator*, we replace their tokens with a special token `[UNK]` which means unknown. With regard to *event*, we convert it into a suitable copula, such as "am" and "is", to ensure the integrity of syntax. Furthermore, we also set up a baseline by randomly replacing a token with `[UNK]` in the context. Examples corresponding to all settings are shown in Table 8 in the Appendix B. We finally report the model accuracy and declined absolute accuracy after hiding the information of each component.

4.2.2 Results

The results in Table 4 show varying degrees of the decline of all settings. If the model's performance decreases more, it means that the influence of the component is more significant than others. Three

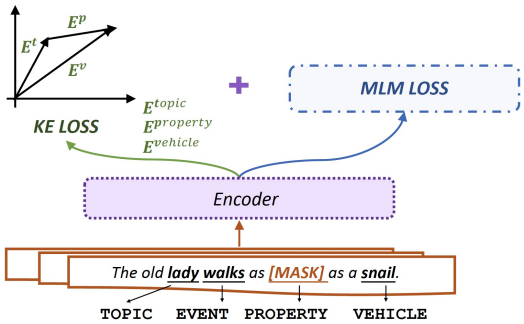

Figure 4: An overview of our objective function design

major components (i.e., vehicle, topic and comparator) obtain higher declined absolute accuracy than random token, which demonstrates that the information of these simile components is more valuable than other words to infer the shared properties. Among all the components, removing the *comparator* may cause the most significant performance drop. This result is mostly because PLMs cannot identify the sentence as a simile without an obvious indicator. When it comes to the remaining 3 components, *vehicle* contributes the most, followed by *topic*. Hence, we argue that it may be beneficial to explicitly leverage both the information of *vehicle* and *topic* to infer the properties.

5 Enhancing PLMs with Knowledge

5.1 Knowledge-enhanced Objective

Benefiting from the result that topic and vehicle are the two most essential components for predicting the shared properties of similes, we catch an insight that *property* can be seen as the **relation** between *topic* and *vehicle* following a set of knowledge embedding (KE) methods (Bordes et al., 2013; Wang et al., 2014; Ji et al., 2015).

To integrate the insight mentioned above into our training procedure, we design an objective function as shown in Figure 4. Inspired by triplets representing the relational facts, we can also extract the topic, property, and vehicle from a simile as a triplet (t, p, v). The distance between topic and vehicle in the embedding space represents the plausibility of property. The plausibility can be measured by scoring functions (Bordes et al., 2013; Wang et al., 2014; Ji et al., 2015). To this end, we follow the scoring function from TransE (Bordes et al., 2013) and define the following Mean Square Error (MSE) loss as our KE loss:

$$\mathcal{L}_{\text{KE}} = \text{MSE}(E^t + E^p, E^v) \quad (1)$$

Datasets	Models	Topic	Vehicle	Event	Comparator	Random
General Corpus	**BERT$_{\text{BASE}}$**	59.87*(-04.26)*	54.58*(-09.55)*	62.84*(-01.29)*	46.32*(-17.81)*	63.05*(-01.08)*
	BERT$_{\text{LARGE}}$	67.74*(-04.65)*	61.16*(-11.23)*	70.19*(-02.20)*	46.06*(-26.33)*	69.07*(-03.32)*
	RoBERTa$_{\text{BASE}}$	65.29*(-04.26)*	61.03*(-08.52)*	68.52*(-01.03)*	50.32*(-19.23)*	67.31*(-02.24)*
	RoBERTa$_{\text{LARGE}}$	76.90*(-02.07)*	69.68*(-09.29)*	77.55*(-01.42)*	54.97(-24.00)	77.72*(-01.25)*
Quizzes	**BERT$_{\text{BASE}}$**	67.02*(-07.34)*	62.35*(-12.01)*	73.43*(-00.93)*	52.80*(-21.56)*	71.91*(-02.45)*
	BERT$_{\text{LARGE}}$	77.86*(-05.36)*	64.57*(-18.65)*	82.63*(-00.59)*	55.24*(-27.98)*	79.91*(-03.31)*
	RoBERTa$_{\text{BASE}}$	76.11*(-06.76)*	69.00*(-13.87)*	81.47*(-01.40)*	55.24*(-27.63)*	77.58*(-05.29)*
	RoBERTa$_{\text{LARGE}}$	83.80*(-03.61)*	74.24*(-13.17)*	86.60*(-00.81)*	60.84*(-26.57)*	85.12*(-02.29)*

Table 4: Accuracy of PLMs in the zero-shot setting before and after hiding the information of each component on two datasets.

Datasets	Models	$\mathcal{L}_{\text{MLM}}$	$\mathcal{L}_{\text{Ours}}$	Gain
General Corpus	**BERT$_{\text{BASE}}$**	67.74	69.25	+1.51
	BERT$_{\text{LARGE}}$	73.85	74.07	+0.22
	RoBERTa$_{\text{BASE}}$	70.58	71.74	+1.16
	RoBERTa$_{\text{LARGE}}$	78.97	78.97	+0.00
Quizzes	**BERT$_{\text{BASE}}$**	82.05	82.94	+0.89
	BERT$_{\text{LARGE}}$	84.58	85.94	+1.36
	RoBERTa$_{\text{BASE}}$	84.69	84.89	+0.20
	RoBERTa$_{\text{LARGE}}$	88.97	89.40	+0.43

Table 5: Accuracy of PLMs using MLM and our objectives in our probing task.

Models	Original	$\mathcal{L}_{\text{MLM}}$	$\mathcal{L}_{\text{Ours}}$
BERT$_{\text{BASE}}$	84.96	85.45	**85.63**
BERT$_{\text{LARGE}}$	86.02	86.65	**86.95**
RoBERTa$_{\text{BASE}}$	88.51	88.61	**89.51**
RoBERTa$_{\text{LARGE}}$	88.84	89.08	**90.21**

Table 6: Accuracy of PLMs with three settings in the downstream task of sentiment classification.

where E^t, E^p, E^v are the representations of topic, property and vehicle encoded by PLMs. We also try more advanced methods such as TransH (Wang et al., 2014) and TransD (Ji et al., 2015) for the knowledge embedding objective, and their results are presented in Table 7 in the Appendix A.2.

Finally, our training procedure is to optimize MLM loss and KE loss jointly:

$$\mathcal{L}_{\text{Ours}} = \alpha \mathcal{L}_{\text{KE}} + \mathcal{L}_{\text{MLM}} \quad (2)$$

where α is a hyperparameter used to balance two objective functions.

5.2 Results

Table 5 presents the performance of the models fine-tuned with the MLM objective and our knowledge-enhanced objective on the two datasets, where the last column shows the performance gains brought by our improvement to the training objective. Overall, each model trained with our knowledge-enhanced objective outperforms the one trained with the MLM objective on both datasets, demonstrating the effectiveness of our objective in the probing task.

For the Quizzes dataset, BERT achieves more performance gains than RoBERTa does, which is probably because RoBERTa has better modeled the relationship among *topic*, *property* and *vehicle* in the similes with simple syntactic structure during fine-tuning with the MLM objective. For the General Corpus dataset, the BASE version of models tends to yield higher performance improvements, probably because the models with larger parameter sizes can better capture the relationship among simile components in the similes with more diverse contexts when fine-tuning with the MLM objective.

5.3 Experiments with Downstream Tasks

Similes generally transmit a positive or negative view due to the shared properties (Fishelov, 2007; Li et al., 2012; Qadir et al., 2015). Taking the simile *"the lawyer is like a shark"* as an example, the implicit shared property *"aggressive"* between *"lawyer"* and *"shark"* indicates the negative polarity. Therefore, we design a sentiment polarity downstream task to validate the improvement of our method to infer shared properties.

Our experiments are based on the Amazon reviews dataset[8] which provides reviews and their corresponding sentiment ratings. Following (Mudinas et al., 2012; Haque et al., 2018), we first process the dataset into a binary sentiment classification task by defining the 1-star and 2-star ratings as negative, the 4-star, and 5-star ratings as positive, while excluding the 3-star neutral ratings. To further address the label imbalance problem, we then sample the positive and negative reviews at 1:1. The final dataset consists of 5,023 reviews and is split into a ratio of 6:2:2 for the train/dev/test set.

When performing the sentiment classification task, we only update the parameters of the multi-

[8]https://www.kaggle.com/bittlingmayer/amazonreviews

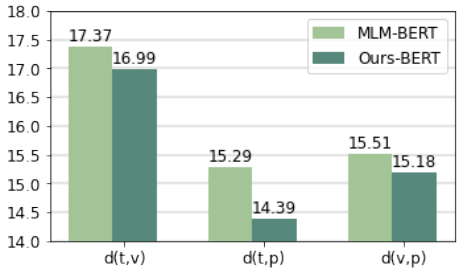

Figure 5: The average semantic distances between the representations of topic(t), property(p), and vehicle(v) in the last layer's hidden state given by $\text{BERT}_{\text{BASE}}$ with MLM and our objectives.

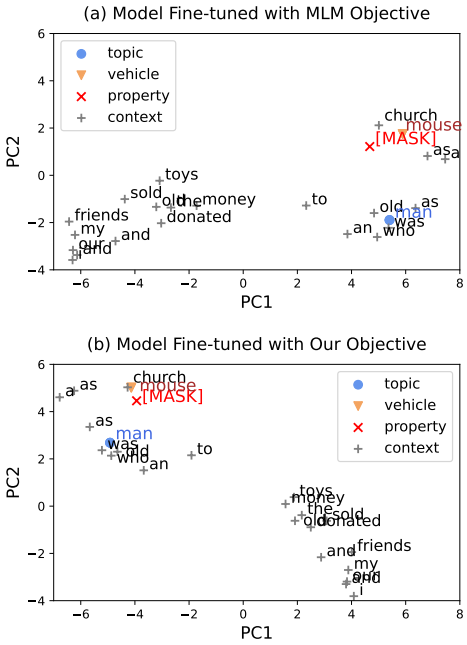

Figure 6: PCA representations of tokens in the last layer's hidden state given by $\text{BERT}_{\text{BASE}}$ with MLM and our objectives.

layer perceptron (MLP) classifiers on top of PLM's contextualized representation. The parameters of PLM are fixed and from three settings: (1) zero-shot; (2) fine-tuned with the MLM objective in the probing task; (3) fine-tuned with the knowledge-enhanced objective in the probing task. The results are shown in the Table 6. First of all, fine-tuning with the MLM objective improves the performance of all models in the sentiment classification task, demonstrating that improving models' ability to infer the properties of similes can enhance models' understanding of the sentiment polarity. Moreover, the performance is further improved by our knowledge-enhanced objective, especially for RoBERTa whose main gains are mostly contributed by our additional knowledge embedding objective. This indicates the effectiveness of our knowledge-enhanced objective in the downstream task of sentiment analysis.

5.4 Analysis

Furthermore, we investigate the mechanism of how knowledge-enhanced objective brings improvement. We first calculate the L2 distance between the representations in the last hidden states of each pair of components. The results are shown in Figure 5. In all pairs, the distance given by our objective is generally shorter than MLM-BERT, which indicates that modeling the relationships among the three important components is efficient to enhance the model performance.

Specifically, we visualize the final layer representation of a simile into two-dimensional spaces via Principal Component Analysis (PCA) (Pearson, 1901) in Figure 6. In both MLM and our objective, the models are required to fill in the masked token in the same simile sentence. The model fine-tuned with the MLM objective predicts wrongly, while our fine-tuned model predicts correctly. We find that our representations of the three components are closer to each other.

6 Related Work

Simile Processing. Simile processing mainly involves 3 fields: simile detection, simile generation, and simile interpretation. The bulk of work in similes mainly focuses on identifying similes and their components (Niculae, 2013; Niculae and Danescu-Niculescu-Mizil, 2014; Liu et al., 2018; Zeng et al., 2020). Recent years have witnessed a growth of work to transfer literal sentences to similes (Zhang et al., 2020; Chakrabarty et al., 2020b). (Chakrabarty et al., 2021b) study the ability of PLMs to recognize textual entailment related to similes. With regard to simile interpretation, (Qadir et al., 2016; Xiao et al., 2016; Bar et al., 2020; Zheng et al., 2019) rank the properties by the statistical co-occurrence and embedding similarities with other simile components. (Chakrabarty et al., 2021a) interpret similes by choosing or generating continuation for narratives via PLMs. Different from these works, we investigate the ability of PLMs to infer shared properties of similes.

Probing Tasks for PLMs. Many studies investigate whether PLMs encode knowledge in their contextual representations by designing probing tasks. Early studies mainly focus on the linguistic knowledge captured by PLMs (Liu et al., 2019a; Tenney et al., 2019). (Petroni et al., 2019) first propose a word prediction task to probe factual

knowledge stored in PLMs. Similar methods are utilized to evaluate various commonsense knowledge, such as symbolic reasoning ability (Talmor et al., 2020; Zhou et al., 2020), numerical commonsense knowledge (Lin et al., 2020), properties associated with concepts (Weir et al., 2020). To our best knowledge, we are the first work to investigate the ability of PLMs to interpret similes by proposing a simile property probing task.

Enhance PLMs via Knowledge Regularization. Recently, many researchers integrate external knowledge with PLMs by complementing the MLM objective with an auxiliary knowledge-based objective. For example, there are works that introduce span-boundary objective for span-level prediction (Joshi et al., 2020), copy-based training objective for mention reference prediction (Ye et al., 2020), knowledge embedding objective for factual knowledge (Wang et al., 2021) and arithmetic relationships of linguistic units for universal language representation (Li and Zhao, 2021). Different from these works, we incorporate simile knowledge with the training objective by modeling the relationship between the salient components of similes.

7 Conclusion

In this work, we are the first to investigate the ability of PLMs in simile interpretation via a proposed novel simile property probing task. We construct two multi-choice probing datasets covering two data sources. By conducting a series of empirical experiments, we prove that PLMs exhibit the ability to infer simile properties in the pre-training stage and further induce more related knowledge during the fine-tuning stage, but there is still a gap between PLMs and humans in this task. Furthermore, we propose a knowledge-enhanced training objective to bridge the gap, which shows effectiveness in the probing task and the downstream task of sentiment classification. In future work, we are interested in exploring the interpretation of more sophisticated figurative language, such as metaphor or analogy.

Acknowledgements

We would like to thank anonymous reviews for their helpful comments and suggestions. Also, thanks to Jingping Liu, Leyang Cui for their insightful feedback that helped improve the paper. We also thank Botian Jiang, Shuang Li for supporting our data collection. This research was supported by the National Key Research and Development Project (No. 2020AAA0109302), National Natural Science Foundation of China (No. 62072323), Shanghai Science and Technology Innovation Action Plan (No. 19511120400), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0103).

Ethical Consideration

We provide details of our work to address potential ethical concerns. In our work, we propose a simile property probing task and construct probing datasets from both general textual corpora and human-designed questions. First of all, all the data sources used in the data collection process are publicly available. Specifically, we follow the robots.txt[9] to respect the copyright when we collect similes from the learning platform Quizizz (Sec. 3.2.1). Moreover, there are three steps involving human annotation to ensure the quality of the datasets: simile and simile components recognition (Sec. 3.2.1), human confirmation for distractors (Sec. 3.2.2), and human performance (Sec. 4.1). To ensure the quality of annotation, all the annotators do not participate in our probing data collection, and they always label a small set of 50 examples to reach an agreement on the labeling criteria before the formal labeling. We protect the privacy rights of annotators and pay them above the local minimum wage.

References

Kfir Bar, Nachum Dershowitz, and Lena Dankin. 2020. Automatic metaphor interpretation using word embeddings. arXiv preprint arXiv:2010.02665.

Monroe C Beardsley. 1981. Aesthetics, problems in the philosophy of criticism. Hackett Publishing.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 26.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. arXiv preprint arXiv:1906.05317.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2021a. It's not rocket science: Interpreting figurative language in narratives. arXiv preprint arXiv:2109.00087.

[9] https://quizizz.com/robots.txt

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021b. Figurative language in recognizing textual entailment. arXiv preprint arXiv:2106.01195.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020a. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6455–6469, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020b. Generating similes< effortlessly> like a pro: A style transfer approach for simile generation. arXiv preprint arXiv:2009.08942.

Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2021. On commonsense cues in bert for solving commonsense tasks. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 683–693.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

David Fishelov. 2007. Shall i compare thee? simile understanding and semantic categories.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. Entropy, 22(1):126.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. arXiv preprint arXiv:1901.05287.

Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education, 15(3):309–333.

Patrick Hanks. 2005. Similes and sets: The english preposition like. Languages and Linguistics: Festschrift for Fr. Cermak. Charles University, Prague.

Patrick Hanks. 2013. Lexical analysis: Norms and exploitations. Mit Press.

Tanjim Ul Haque, Nudrat Nawal Saber, and Faisal Muhammad Shah. 2018. Sentiment analysis on large scale amazon product reviews. In 2018 IEEE international conference on innovative research and development (ICIRD), pages 1–6. IEEE.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 687–696.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8:64–77.

Joyca PW Lacroix, Jaap MJ Murre, and Eric O Postma. 2005. Interpretive diversity as a source of metaphor-simile distinction. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 27.

Bin Li, Haibo Kuang, Yingjie Zhang, Jiajun Chen, and Xuri Tang. 2012. Using similes to extract basic sentiments across languages. In International Conference on Web Information Systems and Mining, pages 536–542. Springer.

Yian Li and Hai Zhao. 2021. Pre-training universal language representation. arXiv preprint arXiv:2105.14478.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. arXiv preprint arXiv:1909.02151.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. arXiv preprint arXiv:2005.00683.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. BT technology journal, 22(4):211–226.

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1543–1553.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. arXiv preprint arXiv:1903.08855.

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. arXiv preprint arXiv:1907.11692.

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

Andrius Mudinas, Dell Zhang, and Mark Levene. 2012. Combining lexicon and learning based approaches for concept-level sentiment analysis. In Proceedings of the first international workshop on issues of sentiment discovery and opinion mining, pages 1–8.

Vlad Niculae. 2013. Comparison pattern matching and creative simile recognition. In Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora, pages 110–114.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 2008–2018.

Vlad Niculae and Victoria Yaneva. 2013. Computational considerations of comparisons and similes. In 51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop, pages 89–95.

Anthony M Paul. 1970. Figurative language. Philosophy & Rhetoric, pages 225–248.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science, 2(11):559–572.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? arXiv preprint arXiv:1909.01066.

Ashequl Qadir, Ellen Riloff, and Marilyn Walker. 2015. Learning to recognize affective polarity in similes. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 190–200.

Ashequl Qadir, Ellen Riloff, and Marilyn Walker. 2016. Automatically inferring implicit properties in similes. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1223–1232.

Siyu Ren and Kenny Q Zhu. 2020. Knowledge-driven distractor generation for cloze-style multiple choice questions. arXiv preprint arXiv:2004.09853.

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, 8:743–758.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. arXiv preprint arXiv:1905.06316.

Tony Veale and Yanfen Hao. 2007. Learning to understand figurative language: From similes to metaphors to irony. In Proceedings of the annual meeting of the cognitive science society, volume 29.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. Transactions of the Association for Computational Linguistics, 9:176–194.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 28.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. arXiv preprint arXiv:2004.04877.

Ping Xiao, Khalid Alnajjar, Mark Granroth-Wilding, Kat Agres, Hannu Toivonen, et al. 2016. Meta4meaning: Automatic metaphor interpretation using corpus-derived word associations. In Proceedings of the Seventh International Conference on Computational Creativity. Sony CSL Paris.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. arXiv preprint arXiv:2004.06870.

Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2020. Neural simile recognition with cyclic multitask learning and local attention. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9515–9522.

Jiayi Zhang, Zhi Cui, Xiaoqiang Xia, Yalong Guo, Yanran Li, Chen Wei, and Jianwei Cui. 2020. Writing polishment with simile: Task, dataset and a neural approach. arXiv preprint arXiv:2012.08117.

Danning Zheng, Ruihua Song, Tianran Hu, Hao Fu, and Jin Zhou. 2019. "love is as complex as math": Metaphor generation system for social chatbot. In Workshop on Chinese Lexical Semantics, pages 337–347. Springer.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9733–9740.

A Additional Experimental Results

A.1 Performance on Different Categories

We investigate whether PLMs are better at inferring the properties of certain categories. Figure 7 presents the performance of the strongest version from each group of models for each category in the zero-shot setting. We found that models perform significantly well when inferring the color, which is probably because each object often has a specific color which in many cases can be inferred without context. However, when it comes to the properties requiring an understanding of the context, such as the personality and qualities *(intelligent, brave)*, temporal properties *(ancient, swift)* and short-term state *(busy, safe)*, models tend to have relatively lower accuracy.

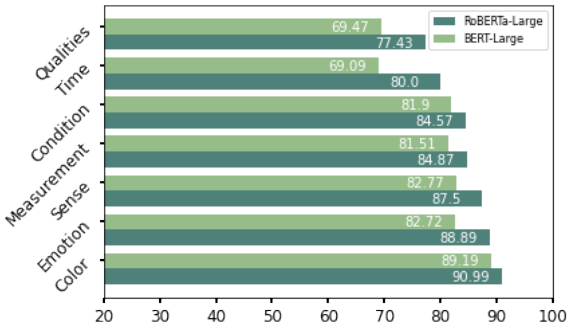

Figure 7: The average accuracy for each category in the zero-shot setting. We select the strongest version from each group of models.

A.2 Comparison of Knowledge Embedding Methods

We also exploit the effects of different knowledge embedding methods when designing our knowledge-enhanced objective. Table 7 shows the performance given by the objectives applying different knowledge embedding methods. First of all, complementing the MLM objective with our knowledge embedding methods generally improves the performance, demonstrating the effectiveness of our approach to enhancing PLMs with simile knowledge. Moreover, following the scoring function from TransE (Bordes et al., 2013) brings the best result in most cases, which indicates that the knowledge embedding methods of simple design are sufficient to incorporate simile knowledge into PLMs in our objective design.

Datasets	Models	$\mathcal{L}_{\text{MLM}}$	$\mathcal{L}_{\text{Ours}}$	$\mathcal{L}_{\text{TransH}}$	$\mathcal{L}_{\text{TransD}}$
General Corpus	$\text{BERT}_{\text{BASE}}$	67.74	69.25	**69.72**	68.38
	$\text{BERT}_{\text{LARGE}}$	73.85	74.07	**74.33**	73.85
	$\text{RoBERTa}_{\text{BASE}}$	70.58	**71.74**	71.18	70.97
	$\text{RoBERTa}_{\text{LARGE}}$	78.97	78.97	78.97	78.97
Quizzes	$\text{BERT}_{\text{BASE}}$	82.05	**82.94**	82.25	82.05
	$\text{BERT}_{\text{LARGE}}$	84.58	**85.94**	85.24	84.69
	$\text{RoBERTa}_{\text{BASE}}$	84.69	**84.89**	84.81	84.81
	$\text{RoBERTa}_{\text{LARGE}}$	88.97	**89.40**	89.32	88.96

Table 7: Comparison of different knowledge embedding methods when designing the knowledge-enhanced objective in our probing task.

B Experimental Details

We introduce details about the implementation of our experiments. The implementations of all the PLMs in our paper are based on the HuggingFace Transformers[10]. During fine-tuning for the probing task, the experiments are run with batch sizes in {8, 16}, α in {3, 5, 10}, a max sequence length of 128, and a learning rate of 1e-5 for 10 epochs. For each model, we use the same hyper-parameters when applying different training objectives. During fine-tuning for the sentiment analysis task, we only update the parameters of the multi-layer perceptron (MLP) classifiers on top of PLM's contextualized representation. We set the learning rate in {2e-5, 3e-5, 4e-5}, batch size of 32, max sequence length of 128 and train for 200 epochs. Additionally, we present examples of the experimental setup for evaluating the influence of important components in Table 8.

Component	Sentence Example
Original	Johan runs as `[MASK]` as a deer to the toilet after he had some spicy gravy .
Topic	`[UNK]` runs as `[MASK]` as a deer to the toilet after he had some spicy gravy .
Vehicle	Johan runs as `[MASK]` as `[UNK]` to the toilet after he had some spicy gravy .
Event	Johan is as `[MASK]` as a deer to the toilet after he had some spicy gravy .
Comparator	Johan runs `[UNK]` `[MASK]` `[UNK]` a deer to the toilet after he had some spicy gravy .
Random	Johan runs as `[MASK]` as a deer `[UNK]` the toilet after he had some spicy gravy .

Table 8: Examples of experiment set-up for evaluating the influence of important components.

C Dataset Description

We introduce details about our classification of the categories of properties. We ask two annotators to label the category of each property in the given sentence and ensure that they agree on the questions that they gave completely different annotation

[10]https://github.com/huggingface/transformers/

Category	Property Example	%
Qualities	strong, weak, cruel, intelligent, brave	27.78
Condition	bad, busy, idle, safe, vain	22.28
Sense	cold, warm, bitter, soft, loud	17.20
Measurement	big, scarce, numerous, tall, tiny	14.16
Color	red, black, green, white, blue	06.75
Time	ancient, new, swift, slow, regular	06.57
Emotion	excited, angry, sad, mad, nervous	05.26

Table 9: Percentage and examples of each category of properties in constructed simile property probing datasets.

results. Table 9 shows the percentage and five examples for each category (possibly more than one category per property). In particular, properties in *Quialities* describe the long-term feature of a material or a person's character, while properties in *Condition* depict a short-term state. Table 1 presents the percentage and examples for our simile probes of different categories.