Towards Semantics-Enhanced Pre-Training: Can Lexicon Definitions Help Learning Sentence Meanings?

Xuancheng Ren,[1] **Xu Sun,**[1,2*] **Houfeng Wang,**[1] **Qun Liu**[3]
[1] MOE Key Laboratory of Computational Linguistics, School of EECS, Peking University
[2] Center for Data Science, Peking University
[3] Huawei Noah's Ark Lab
{renxc, xusun, wanghf}@pku.edu.cn, qun.liu@huawei.com

Abstract

Self-supervised pre-training techniques, albeit relying on large amounts of text, have enabled rapid growth in learning language representations for natural language understanding. However, as radically empirical models on sentences, they are subject to the input data distribution, inevitably incorporating data bias and reporting bias, which may lead to inaccurate understanding of sentences. To address this problem, we propose to adopt a human learner's approach: when we cannot make sense of a word in a sentence, we often consult the dictionary for specific meanings; but can the same work for empirical models? In this work, we try to inform the pre-trained masked language models of word meanings for semantics-enhanced pre-training. To achieve a contrastive and holistic view of word meanings, a definition pair of two related words is presented to the masked language model such that the model can better associate a word with its crucial semantic features. Both intrinsic and extrinsic evaluations validate the proposed approach on semantics-orientated tasks, with an almost negligible increase of training data.

Introduction

Pre-trained language representations have evolved remarkably in recent years, from static word representations (Mikolov et al. 2013; Pennington, Socher, and Manning 2014) to contextual word representations (Peters et al. 2018; Radford et al. 2018; Howard and Ruder 2018; Devlin et al. 2019; Yang et al. 2019; Brown et al. 2020). Typically, they are first pre-trained on large-scale unannotated data and then used in downstream tasks. As computational realization of distributional semantics, where the meaning of a word can be decided from its linguistic context, they have proved highly effective and produced state-of-the-art human-level performances on several datasets in natural language understanding (Wang et al. 2019b).

However, recently studies have suggested that while pre-trained models are adequate in capturing the syntax of language (Hewitt and Manning 2019; Clark et al. 2019), they often produce inaccurate interpretation of a sentence (Niven and Kao 2019; McCoy, Pavlick, and Linzen 2019). We collect several representative cases in Table 1. As we can see, the model seems to prefer the safer choices that are more frequent in related context, neglecting the semantics of the sentence or the low frequency words in the sentence.

People usually eat the [M] of the watermelon. $p(\text{seeds}) = 0.20 > p(\text{flesh}) = 0.06$
Drinking too much [M] can make people drunk. $p(\text{water}) = 0.24 > p(\text{alcohol}) = 0.17$
A ratchet is used for moving in [M] direction. $p(\text{any}) = 0.36 > p(\text{one}) = 0.34$
A scapegrace is an [M] rascal. $p(\text{innocent}) = 0.11 > p(\text{incompetent}) = 0.03$

Table 1: Masked words recovered by RoBERTa-large. Between the two syntactically plausible replacements, the model prefers the semantically insensible one, which suggests that such models may not really understand the semantics of the sentence but rely more on distributional statistics.

This phenomenon could root from the entirely empirical approach adopted by BERT-like models. In theory, the meaning of a word can be decided from all its use according to the hypothesis of distributional semantics (Wittgenstein 1953; Harris 1954; Weaver 1955; Firth 1957; Boleda 2020). However, in practice, the problems are three-fold: (a) word contexts are sampled, since the use of a word is inexhaustible, which means the learning is inaccurate and artefacts in data will be exploited (Gururangan et al. 2018; Niven and Kao 2019; McCoy, Pavlick, and Linzen 2019), which can be remedied in part by enlarging the data scale (Liu et al. 2019; Yang et al. 2019; Raffel et al. 2020); (b) different word types are not equally represented in data, which means words of lower frequency are harder to learn in whatever data scale; and more importantly (c) certain semantics are not naturally expressed in sentences due to reporting bias, which is a problem that cannot be addressed by empirical approaches (Lucy and Gauthier 2017; Da and Kasai 2019; Forbes, Holtzman, and Choi 2019; Kwon et al. 2019).

In contrast, words, which are basic semantic units of sentences, are countable and are equally represented in lexicons. Now that the pre-trained models can encode the sentence structure well, is it possible to improve the word coverage and correct the misconceptions of the model by directly

*Corresponding Author

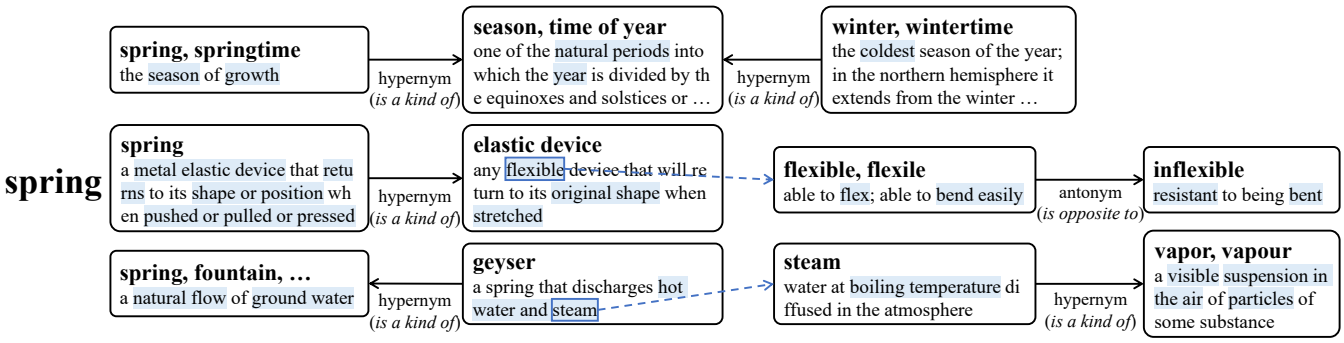


Figure 1: Examples for the word *spring* from WordNet organized in synsets with definitions. The synsets are linked by semantic relations. As *spring* is a polysemous word, it corresponds to multiple synsets. As we can see, the definitions contain rich semantic features (colored) for a word and relations can further broaden the depth of the semantic features.

injecting word meanings? Just as when we read a sentence containing an unfamiliar word, as long as we know the word itself better, we can understand the sentence without reading hundreds or thousands of sentences with the new word.

In this paper, we explore directly injecting word meanings into pre-trained models based on *definitional* lexical semantic knowledge in external linguistic knowledge bases, as shown in Figure 1. We devise two complementary tasks to absorb such knowledge, which are predicting the word given its definition and recovering certain parts of the definition with the word. The two tasks are united as the task of masked word-definition prediction. Considering the relational knowledge between words, we strengthen the depth of word meaning by pairing related word-definition sequences. The injection of explicit lexical semantics broadens the lexicon coverage and semantic features encoded in the pre-trained models with almost negligible new training data, i.e., 1000x times fewer than the data used by the baseline, and both intrinsic and extrinsic evaluations show promising results on semantics-oriented tasks.

In all, our contributions are summarized as follows:

- We explore enhancing the semantic capabilities of pre-trained masked language models with lexical semantics in terms of word definitions to remedy the limited coverage and possible misconceptions of the model.
- The objective is cast as a masked word-definition task that injects explicit definitional knowledge containing rich semantic features. Such knowledge is propagated on the semantic network to provide a contrastive and comprehensive view of word meanings.
- The proposed approach is simple yet effective and can help the model to learn better semantic representations, which is verified by intrinsic and extrinsic evaluations.

Injecting Word Meanings

The masked language model objective is the key to self-supervised pre-training but requires large quantities of data to learn reasonable representations. However, they still face challenges in comprehensive understanding of words and the related semantics. Drawing inspirations from human's learning practice, we explore dealing with the challenges by directly incorporating knowledge of words. If the pre-trained models have a strong sense of how the sentences are constructed, as suggested by many previous studies (Clark et al. 2019; Tenney et al. 2019; Goldberg 2019; Niven and Kao 2019; Reif et al. 2019; Warstadt et al. 2019; Lin, Tan, and Frank 2019; Hewitt and Manning 2019), it should be possible to inform the models with the meaning of words, which in turn streamlines the deduction of the meaning of sentences. It is also desirable that the introduction of word meanings remains supplementary to the empirical data and compatible with the original approaches so that we can combine the best of both worlds.

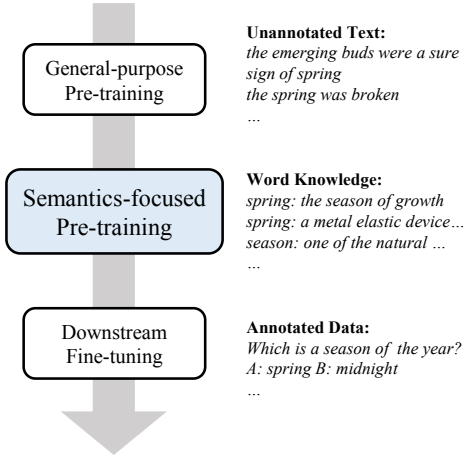
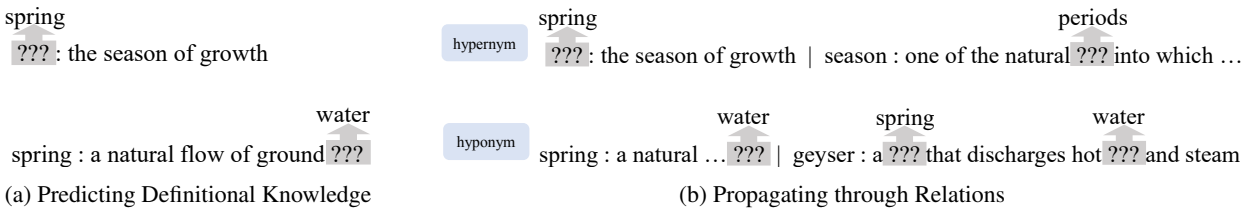


Figure 2: Illustration of the training process. The general-purpose pre-training builds the model's ability to understand sentences on unannotated text; the semantics-focused pre-training injects word knowledge to the model to increase its coverage and rectify misreadings with external linguistic knowledge bases; then the pre-trained model is applied to downstream tasks for task-specific fine-tuning.

The proposed approach consists of three consecutive training stages, as illustrated in Figure 2: (1) general-purpose pre-training, where the model is shown large-scale unan-



(a) Predicting Definitional Knowledge (b) Propagating through Relations

Figure 3: Examples of the proposed semantics-focused pre-training. We draw word knowledge from classical word definitions as in Figure 3(a), which contain rich semantic features, and broaden the scope by relations to other words as in Figure 3(b).

notated text to learn to interpret sentences, (2) semantics-focused pre-training, where the model is enhanced in semantic capabilities with the help of definitional and relational word knowledge, and (3) downstream fine-tuning, where the model adapts to the target task.

General-Purpose Pre-Training

General-purpose pre-training follows the existing successful self-supervised technique, specifically using the masked language model objective, and aims to understand the structure of language with large-scale unannotated text (Clark et al. 2019; Tenney et al. 2019). As syntax learning and semantics learning are highly coupled, this stage also promotes the model in some semantic aspects, e.g., selectional preferences (Ettinger 2020; Tenney et al. 2019), which have prominent syntactic clues.

The input to BERT-like masked language models (Devlin et al. 2019; Liu et al. 2019) is a sequence of word tokens $\mathbf{x} = \{x_i\}_{i=1}^{N}$, where N is the input sequence length. 15% of the tokens are further masked, of which 80% are replaced with the *[M]* token, 10% are replaced with a random token, and 10% are kept unchanged. Let $\hat{\mathbf{x}}$ denote the sequence after masking. A neural network, most commonly a stack of Transformer encoder blocks (Vaswani et al. 2017), is applied to the input sequence to obtain the contextualized representation $\mathbf{s}_i$ for each word token. For each masked position j, a classification network is tasked to predict the original word token using cross-entropy loss over the entire vocabulary $\mathcal{D}$:

$$\mathcal{L} = -\log p(x_j|\hat{\mathbf{x}}), \tag{1}$$

where x_j is the original word token before masking.

Semantics-Focused Pre-Training

*Sem*antics-focused *pre*-training (**SemPre**) aims to enhance the learning of sentence meaning by learning word meanings. However, the fundamental question is: what is the meaning of a word and how can it be described? Moreover, can word meanings stand alone aside from context or use and is there a borderline between linguistic knowledge and encyclopedic knowledge? Those topics are extensively discussed in the literature of generative lexical semantics and a number of approaches to represent word meanings have been proposed, such as lexical field theory, componential analysis, and relational semantics. Our investigation builds on a practical, mixed view of those theories.

To be precise, we rely on the classical dictionary-like definitions for semantic features and the relational structures among the words for a conceptual understanding of word meanings. Such lexical knowledge is drawn from WordNet (Miller 1995) and some examples are shown in Figure 1. WordNet is organized in synonym sets, also known as synsets, which group concrete "words" (including multi-word expressions, e.g., *hot dog*) of similar meanings together. A word can belong to multiple synsets, in which case polysemy occurs. Each synset is accompanied with a definition, example sentences, and semantic relations to other synsets. As the natural input to pre-trained language models is words instead of synsets, we flatten synsets to words and introduce a synonym relation among the words in a synset to account for the synset information.

Predicting Definitional Knowledge The definition of a word contains rich semantic features, which are curated, noise-free, and deemed essential by lexicographers. In order to familiarize the model with such features, we recast the masked language model task in the previous training stage to operate on masked word-definition sequences. For a word w and its definition sentence sequence $\mathbf{d} = \{d\}_{i=1}^{M}$ in the lexicon $\mathcal{A}$, an artificial sentence is constructed in the format of $\mathbf{x}^{(w)} = \{w, \texttt{[D]}, d_1, d_2, \ldots, d_M\}$, where M is the length of the definition and *[D]* is a special symbol separating the word and the definition to allow for multi-token words. This input format is used because it is non-trivial to construct a grammatically correct sentence and format indicators, in our case *[D]*, can adequately inform the model of structural meanings (Lewis et al. 2020; Raffel et al. 2020). The same masking procedure as general-purpose pre-training is applied and the model is supposed to recover the masked tokens. It corresponds to two special cases. Let's take the definition of *spring* in the season sense as an example:

- "spring [D] the season of [M]": This is a semantic feature filling task. Note that different from real sentences, the definition itself provides no clue of the correct word. The model needs to encode such features in its parameters.
- "[M] [D] the season of growth": This is a word guessing task. As definitions are information-dense, an incomplete reading can lead to wrong prediction. The model needs to associate the features comprehensively.

Propagating Relational Knowledge The classical definition is by no way an exhaustive description of the meaning of a word. Definitions are often written in a way to distinguish one word from similar words. The notion is prominently reflected in the structural view of lexical semantics, which believes the word meaning can be defined by relations among

words in the same semantic field. In light of this, we expand the semantic scope of a word to its related words, whose definitions also contain valuable semantic features. For a relation triple (w_s, r, w_e), we first construct the word-definition sequences $\mathbf{x}^{(s)}$ and $\mathbf{x}^{(e)}$ for the pair of words. Then, the pair of word-definition sequences are connected with a separator as a new sequence. However, it is important to interpret semantic features from other words cautiously, because relations also vary in semantics. For example,

- *spring* and *season* have the hypernym relation and the semantic features of *season* usually applies to *spring*; but
- *hot* and *cold* have the antonym relation and have contradictory semantic features *high* and *low* for temperature.

To this end, we add a classification objective to predict the type of the semantic relation, e.g., hypernym, antonym, etc., from the sequence representations. The input sequence is of the format $\mathbf{x} = \{\texttt{[C]}, x_1^{(s)}, \ldots, x_M^{(s)}, \texttt{[S]}, x_1^{(e)}, \ldots, x_N^{(e)}\}$, where *[C]* and *[S]* are two special tokens, and M and N are the length of the two word-definition sequences, respectively. The representation of the *[C]* token is used as the sequence representation (Devlin et al. 2019) to predict the relation. Let $\hat{\mathbf{x}}$ denote the sequence after masking. The multi-task loss for this stage could be formalized as

$$\mathcal{L}_{\text{sem}} = -\log p(r|\hat{\mathbf{x}}) - \sum\nolimits_{j \in \mathcal{M}} \log p(x_j|\hat{\mathbf{x}}), \qquad (2)$$

where $\mathcal{M}$ is the set of masked positions and we forbid the special tokens to be masked.

With the semantics-focused pre-training, the model is enhanced in its lexicon coverage and injected with explicit semantic features. The approach is conceptually simple and compatible with the previous training stage to avoid catastrophic forgetting of syntactic capability. In addition, it uses far fewer data than the previous training stage.

Experiments

To verify whether the proposed method can enhance the semantic understanding of sentences, we conduct both intrinsic evaluation that inspects knowledge learned by the pre-trained models themselves and extrinsic evaluation on semantics-oriented downstream tasks with fine-tuning.[1]

Tasks

Word Games (WGs) is a word ranking task, where a system is required to predict the word by its definition. The words are extracted from the Oxford 3000 list, which contains the most important English words to learn. Two variants of the datasets are constructed by using definitions from the WordNet (WG1) and the Oxford Advanced Learner's Dictionary (OALD) (WG2). For each pair of a word and its definition, we make up several sentences with various templates to allow the models to understand the sentence better. The evaluation metric is the average of the lowest rank of the target word predicted by the system across the template sentences for an example. For testing, the target word is replaced with the *[M]* token.

Commonsense Ability Tests (CATs) (Zhou et al. 2020) contain eight datasets, i.e., Sense Making (SM), Winograd Schema Challenge (WSC), Conjunction Acceptability (CA), SWAG, HellaSwag, Sense Making with Reasoning (SMR), and Argument Reasoning Comprehension Task (ARCT1 and ARCT2), re-framed from six existing benchmarks (Wang et al. 2019c; Levesque, Davis, and Morgenstern 2012; Zellers et al. 2018, 2019; Habernal et al. 2018; Niven and Kao 2019). The tested system is supposed to determine the sentence that makes sense from several adversarial sentences, which have differences on the token-level or the sentence-level that require various kinds of commonsense knowledge to resolve. The evaluation metric is accuracy. For the testing protocol, we follow Zhou et al. (2020).

Word-in-Context (WiC) is a word sense disambiguation task, where a system is required to determine if the polysemous target word has the same sense in two short sentences with minimum contexts (Pilehvar and Camacho-Collados 2019). For negative cases, the target word senses are ensured to be differentiable and of different supersenses. The evaluation metric is accuracy.

Physical Interaction: Question Answering (PIQA) is a question answering task, where given a goal and two possible solutions, the system must choose the most appropriate solution (Bisk et al. 2020). The dataset is human generated and focuses on everyday situations with physical commonsense knowledge, which is inherently limited in text due to reporting bias, e.g., flexibility and being porous. The evaluation metric is accuracy.

Settings

Our implementation is based on the fairseq (Ott et al. 2019) package. For general-purpose pre-training, we adopt the pre-trained RoBERTa-base and RoBERTa-large models (Liu et al. 2019) to save computation resources, which contain 125M and 355M parameters, respectively. They are trained on a combined corpus including fictions, encyclopedia, and news, totaling over 160GB text, which contains approximately 1500M sentences. The input word cases are kept and the input sequence is pre-processed using BPE. For semantics-focused pre-training, the models are trained on word-definition pairs using the objective in Eq. (2) and we extract 0.2M word-definitions and 1.4M word-definition pairs in 23 relations from WordNet. For the two WG datasets only, we use models that train on parts of WordNet, such that one-third of the words in testing are not seen in training, to avoid memorization that may favor SemPre in WG1. We use a batch size of 2048 sequences, a peak learning rate of 2×10^{-5} with linear warm-up and decay peaked at the 295th update scheduled for at most 6910 updates and keep at most 128 tokens of a sequence. The rest is kept similar to RoBERTa pre-training. We use the model after the first epoch for further evaluations.

For downstream fine-tuning, following Liu et al. (2019); Bisk et al. (2020), we conduct a grid search with respect to certain hyper-parameters, i.e., the learning rates $[1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}]$ and the maximum epochs $[10, 50]$.

[1] For detailed introduction to the tasks and the experimental settings, please refer to the appendix. The code and the appendix are available at https://github.com/lancopku/sempre

Model	Word Game (↓)		Commonsense Ability Tests (↑)							
	WG1	WG2	CA	WSC	SM	SMR	SWAG	HellaSwag	ARCT1	ARCT2
RoBERTa-base (our reproduction)	592	464	95.6	62.5	75.0	40.0	69.1	41.3	50.0	53.7
RoBERTa-base + SemPre	111	236	95.1	63.6	77.7	39.5	68.0	41.7	53.8	55.9
RoBERTa-large (our reproduction)	332	262	96.2	69.3	79.2	47.3	**76.1**	**48.9**	54.3	60.0
RoBERTa-large + SemPre	**86**	**130**	**96.7**	**73.5**	**80.4**	**48.4**	75.9	48.5	**58.3**	**60.9**

Table 2: Intrinsic evaluation without fine-tuning. ↓ denotes the lower the better; ↑ is the opposite. SemPre causes a significant lead in Word Game and enjoys a comfortable margin on CATs, especially on WSC, which highly involves object knowledge rarely reported in natural sentences, except for SWAG and HellaSwag, which involve event and temporal knowledge that is rarely covered by semantic features in definitions.

Model	#Parameters	WiC (↑)		PIQA (↑)	
		Validation	Test	Validation	Test
ELMo-inspired (Ansell, Bravo-Marquez, and Pfahringer 2019)	-	67.4	61.2	-	-
GPT (Bisk et al. 2020)	124M	-	-	70.9	69.2
BERT-large (Wang et al. 2019a; Bisk et al. 2020)	340M	74.9	69.5^{Σ}	67.1	66.8
KnowBERT-W+W (Peters et al. 2019)	523M	72.6	70.9^{Σ}	-	-
RoBERTa-large (Liu et al. 2019; Bisk et al. 2020)	355M	75.6	69.9^{Σ}	79.2	77.1
T5-large (Raffel et al. 2020)	770M	-	69.3	-	-
RoBERTa-base (our reproduction)	125M	69.4	66.9	73.7	-
RoBERTa-base + SemPre	125M	71.5	69.3	75.0	-
RoBERTa-large (our reproduction)	355M	74.6	70.3	81.3	-
RoBERTa-large + SemPre	355M	**75.7**	**72.1**	**81.6**	**79.0**

Table 3: Extrinsic evaluation on downstream tasks with fine-tuning. ↑ denotes the higher the better; $^{\Sigma}$ denotes results obtained using model averaging or model ensemble. Results are directly taken from the related papers. SemPre demonstrates consistent improvements on the tasks relying on various aspects of semantics with a competitive parameter budget.

The batch size is 32. Each configuration is run multiple times with different random start. We adopt early stopping based on validation accuracy and report the results of the best-scoring configuration on the validation set. Results of other models are taken directly from the corresponding papers, except for RoBERTa, for which we also report results of our reproduction for fair comparison.

Results

Intrinsic Evaluation The results on WGs and CATs, which are obtained in a zero-shot manner without fine-tuning, are shown in Table 2. SemPre benefits almost all datasets indicating improved understanding of words and their semantic features.

As we can see, significant improvements are achieved on WGs for definitions from both WordNet and OALD. We report results on all test words. For results on WG1 categorized by whether the word definition is seen in training, please refer to the appendix and the results are similar. One may wonder why the model could predict the word better without training on its definition in SemPre. It is most likely that the models learn to better connect and elicit the semantic features in the sentences, because meanings of words are not isolated and knowledge will transfer to words even if they are unseen in SemPre. For example, knowing *spring* better may help knowing *season* better.

For CATs, SemPre improves the pre-trained models by a comfortable margin except for SWAG and HellaSwag, suggesting that the semantic features in word definitions promote commonsense knowledge to a certain extent, that is, for conceptual and perceptual knowledge tested by CA, WSC, SM, SMR, and ARCT, but not event and temporal knowledge required by SWAG and HellaSwag. Apart from the difference in datasets, model sizes also affect the results and smaller models could be unstable in internalizing some types of knowledge injected by SemPre, as demonstrated by the results of RoBERTa-base models.

Extrinsic Evaluation The results from intrinsic evaluation validate that SemPre can indeed inject lexical knowledge to the pre-trained models, but it is also important to see whether improved learning of lexical semantics can benefit downstream tasks. For this evaluation with fine-tuning, apart from our baseline reproductions, we also compare with other pre-trained models, including ELMo, GPT, BERT, and T5, which are for general domains, and KnowBERT, which enhances BERT with entity embeddings from Wikipedia and WordNet. The results are reported in Table 3.

For WiC, comparing BERT-large and RoBERTa-large, we can see that although RoBERTa-large uses 10 times of data compared to BERT-large, only 0.4 to 0.8 performance increase can be observed, suggesting purely training on more

Model	Word	Def.	Pair	CATs	WiC	Avg
(a) RoBERTa	○	○	○	66.4	74.6	70.5
(b) + Def.	○	●	○	66.9	74.3	70.6
(c) + Word Pair	●	○	●	64.3	74.3	69.3
(d) + Word-Def.	●	●	○	67.6	74.8	71.2
(e) + SemPre	●	●	●	**67.8**	**75.7**	**71.8**
(f) - Word Mask	∗	●	●	67.1	75.1	71.1
(g) - Def. Mask	●	∗	●	67.4	74.9	71.2
(h) - Rel. Pred.	●	●	∗	67.7	74.6	71.1

Table 4: Results of the ablation study. ○, ∗, and ● denote the respective information is not used, used only as inputs, and learned with an objective. We show the average of 8 datasets for CATs. The results on WiC are based on the validation set. Avg denotes the average of CATs and WiC.

data has diminishing benefits on the WiC task, which requires accurate understanding of words. With our approach, a further improvement of 1.8 in accuracy can be obtained.

PIQA targets at physical commonsense in interactive environment, requiring reasoning of affordance (Gibson 1979), e.g., using a sharp kitchen knife to mince a carrot, and specific routine knowledge, e.g., how to put a game in an Xbox. While SemPre can help the former, as physical features in definitions are fundamental for affordance and definitions often link actions and general objects, it can be hard to provide assistance to the latter, which involves named entities and event knowledge. Even so, we still observe a positive lead on PIQA with SemPre, indicating the general usefulness of semantic-focused pre-training. For the validation results using RoBERTa-large, the difference is statistically significant ($p < 0.05$, t-test, one-sided).

We also evaluate on the GLUE benchmark (Wang et al. 2019b) and Semantic Role Labeling on the CoNLL-2012 dataset (Pradhan et al. 2013). Our RoBERTa-base and the SemPre-enhanced version both achieve test scores of 76.4∼76.5 on GLUE and overall F1 measures around 86.9 on CoNLL-2012. For these tasks that have rich contexts and strong syntactic indicators, SemPre maintains the ability of fundamental language understanding. Moreover, we find that those datasets are not effective in evaluating the aspects enhanced by SemPre in terms of coverage in lexical semantics. The detailed results are reported in the appendix.

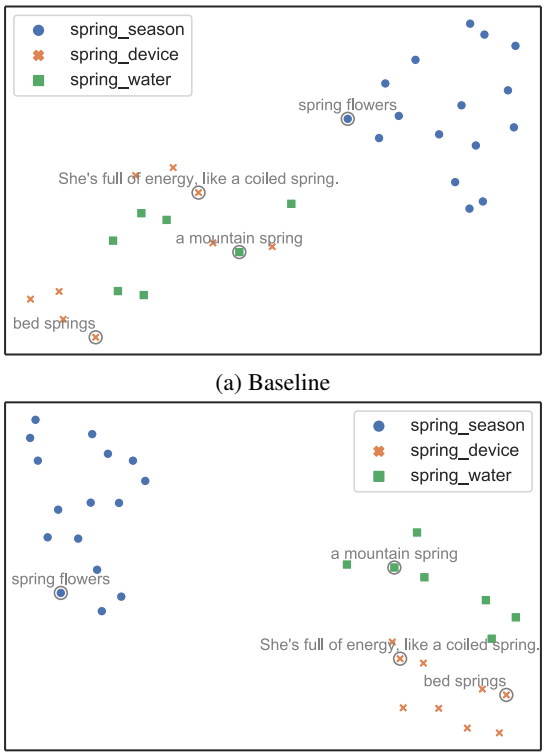
Analysis

Ablation Study As SemPre introduces information sources and techniques to learn such information, the ablation study is conducted to analyze the effects from both perspectives and the results on CATs and WiC using RoBERTa-large are shown in Table 4. Please see the appendix for full CATs results. For ablated models that do not use pairs of words or word-definitions, the training data shrink 7 times and we take the models after the entire training, i.e., 10 epochs, for further evaluation. The upper part of Table 4 studies from the perspective of information sources. As we can see from model (a) to model (e), introducing word-definitions is the main source of improvements and definitions are the crucial component. Moreover, using pairs of words alone as model (c) is not beneficial, and only when pairs of word-definitions are incorporated, that is, the proposed SemPre, the best results can be obtained. The lower part of Table 4 demonstrates the effectiveness of the techniques in harnessing the definitional knowledge. As shown by the results of model (f), model (g), and model (e), masking words and definitions are both instrumental in injecting definitional knowledge. Comparing model (h) and model (d) to model (e), we can see that relation prediction is important when propagating definitions through relations. These results indicate the improvements of SemPre do not come from simply using more data but the way the extra data are used for learning word meanings. We further conduct ablation study in terms of distributional characteristics of words, i.e., part-of-speech, to reveal the systematic tendencies of the pre-trained models that are purely based on language distributions, and the analysis is elaborated in the appendix.

Input: [M] can be used to make furniture. Baseline: It (0.27), They (0.14), This (0.10) + SemPre: Wood (0.68), It (0.16), Steel (0.03)
Input: My dream job is [M]. Baseline: here (0.25), gone (0.07), over (0.06) + SemPre: teaching (0.12), writing (0.05), marketing (0.03)
Input: Fruits are [M]. Baseline: expensive (0.06), important (0.02), optional (0.02) + SemPre: edible (0.36), nutritious (0.05), food (0.02)
Input: Food is for [M]. Baseline: everyone (0.10), people (0.06), you (0.05) + SemPre: eating (0.41), consumption (0.18), living (0.12)

Table 5: Masked sentences recovered by RoBERTa-large and our proposal. We show the top-three predictions and their probabilities. As we can see, the predicted words are more semantics-oriented with SemPre.

Case Study From a qualitative view, we further test whether SemPre helps grasp the meaning of the sentences. The baseline is RoBERTa-large. First, in Table 5, we show some representative cases, where the model is asked to predict the most likely replacements for the masked position. We can see that RoBERTa tends to prefer words that are more general and less specific to the context, while SemPre can improve the prediction in terms of semantic informativeness. Second, in Figure 4, we show the UMAP visualization (McInnes and Healy 2018) of representations for the word *spring* in different contexts, in a similar way to Reif et al. (2019). RoBERTa has difficulty in differentiating the device and the water sense, while with SemPre, a clear boundary can be determined. It is important to note that SemPre enhances the pre-trained models only using the definitions instead of using examples sentences as in Levine et al. (2020), yet it can enhance the comprehension of words in real contexts, suggesting SemPre most likely activates the semantic knowledge in a continual learning manner rather than builds the related skills from scratch.



(a) Baseline

(b) Baseline + SemPre

Figure 4: Visualization for *spring* of different meanings in contexts. Representations generated by SemPre are better clustered in terms of meaning.

Related Work

Self-Supervised Pre-Training This approach typically takes in unannotated data as contexts and predicts certain parts of the input data as the target. It has been used in natural language processing to learn word representations (Mikolov et al. 2013; Pennington, Socher, and Manning 2014) and recently contextualized word representations (Howard and Ruder 2018; Peters et al. 2018). The rather successful BERT-like models (Devlin et al. 2019; Liu et al. 2019) use a masked language model objective and considerable efforts have been devoted to discovering better sequence corrupting scheme (Sun et al. 2019; Joshi et al. 2020; Wang et al. 2020; Clark et al. 2020), applying autoregressive objectives (Song et al. 2019; Dong et al. 2019; Lewis et al. 2020; Yang et al. 2019), reducing the computational complexity (Jiao et al. 2020; Sanh et al. 2019; Sun et al. 2020; Lan et al. 2020), etc. The approach presented in this work can be extended further to those models and is mostly independent to those improvements.

Pre-Training with Semantics Substantial efforts have been put into understanding the BERT-like models and most studies affirm that BERT has learned syntactic structures exceptionally well (Clark et al. 2019; Tenney et al. 2019; Goldberg 2019; Lin, Tan, and Frank 2019; Hewitt and Manning 2019) and while there is evidence that the model encodes certain limited semantic knowledge (Niven and Kao 2019), e.g., selectional preference (Bouraoui, Camacho-Collados, and Schockaert 2020) and semantic categorization (Wiedemann et al. 2019; Ettinger 2020; Reif et al. 2019), such knowledge relies on context providing clues (Bouraoui, Camacho-Collados, and Schockaert 2020). To remedy this, several extensions are proposed to incorporate world knowledge (Sun et al. 2019; Zhang et al. 2019; Peters et al. 2019) or commonsense (Ye et al. 2019; Levine et al. 2020) into BERT, but most of them focused on specific downstream tasks, such as commonsense question answering (Ye et al. 2019) and word sense disambiguation (Loureiro and Jorge 2019; Huang et al. 2019; Levine et al. 2020). Lauscher et al. (2019) incorporated into BERT the linguistic knowledge of the so-called true semantic similarity, consisting of synonym or hypernym-hyponym word pairs, via a binary classification objective. Different from them, we first explore the efficacy of learning word meanings from lexicon definitions.

Semantic Knowledge Sources In this work, we extract definitions and relations from WordNet (Miller 1995). Although it does not contain closed-set words, e.g., determiners and prepositions, its coverage of concrete words is more than adequate for normal language understanding purposes. Moreover, a variety of other resources exist for the application of the proposed method. Definitions via dictionaries are not in shortage, since traditional dictionaries exist for almost all languages and there are wiki-based online dictionaries constantly absorbing new terms, such as Wiktionary and Urban Dictionary. Semantic relations with different focuses are also proposed, including commonsense (Speer, Chin, and Havasi 2017) and sentiment (Baccianella, Esuli, and Sebastiani 2010). Definitions could establish new knowledge sources by mapping to existing semantic relational graphs, which is the practice of Open Multilingual Wordnet (Bond and Paik 2012; Bond and Foster 2013) and BabelNet (Navigli and Ponzetto 2012).

Conclusion and Future Work

We explore a semantics-focused training approach for pre-trained masked language models to see whether better understanding of words can lead to better understanding of sentences. Harnessing the knowledge of an external linguistic knowledge source, both definitional and relational knowledge of lexical semantics are incorporated into the models. In contrast to previous methods that try to instantiate such kinds of knowledge as natural sentences, we make use of a masked word-definition prediction objective with a relation classification objective. The resultant approach is conceptually simple and self-contained without the need of corpora to retrieve related sentences or tools such as entity linkers, but can also elevate the performance on semantics-oriented tasks. Our analysis indicates the proposed intermediate training method drives the model to elicit semantic features from sentences. For future work, we would like to explore more elaborate ways to generate word pairs, especially in terms of abductive relation paths, and pre-training with named entities, e.g., Wikipedia, instead of general words.

Acknowledgments

This work is supported in part by National Key R&D Program of China (No. 2020AAA0105200) and Beijing Academy of Artificial Intelligence (BAAI). We thank all the anonymous reviewers for their constructive comments and Pengcheng Yang, Fuli Luo, and Lei Li for helpful discussion in preparing the manuscript.

References

Ansell, A.; Bravo-Marquez, F.; and Pfahringer, B. 2019. An ELMo-inspired approach to SemDeep-5's Word-in-Context task. In *SemDeep@IJCAI*, 21–25.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *LREC*, 2200–2204.

Bisk, Y.; Zellers, R.; LeBras, R.; Gao, J.; and Choi, Y. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI*, 7432–7439.

Boleda, G. 2020. Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics* 6(1): 213–234.

Bond, F.; and Foster, R. 2013. Linking and Extending an Open Multilingual WordNet. In *ACL (1)*, 1352–1362.

Bond, F.; and Paik, K. 2012. A Survey of WordNets and Their Licenses. In *Global WordNet Conference*, 64–71.

Bouraoui, Z.; Camacho-Collados, J.; and Schockaert, S. 2020. Inducing Relational Knowledge from BERT. In *AAAI*, 7456–7463.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *NeurIPS*, 1877–1901.

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *BlackboxNLP*, 276–286.

Clark, K.; Luong, M.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*.

Da, J.; and Kasai, J. 2019. Cracking the Contextual Commonsense Code: Understanding Commonsense Reasoning Aptitude of Deep Contextual Representations. In *COIN@EMNLP-IJCNLP*, 1–12.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In *NeurIPS*, 13042–13054.

Ettinger, A. 2020. What BERT is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Trans. Assoc. Comput. Linguistics* 8: 34–48.

Firth, J. R. 1957. A Synopsis of Linguistic Theory, 1930-1955. In *Studies in Linguistic Analysis*, 1–32. Blackwell.

Forbes, M.; Holtzman, A.; and Choi, Y. 2019. Do Neural Language Representations Learn Physical Commonsense? In *CogSci*, 1753–1759.

Gibson, J. J. 1979. *The Ecological Approach to Visual Perception*. Houghton Mifflin.

Goldberg, Y. 2019. Assessing BERT's Syntactic Abilities. *CoRR* abs/1901.05287.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL-HLT 2018 (2)*, 107–112.

Habernal, I.; Wachsmuth, H.; Gurevych, I.; and Stein, B. 2018. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. In *NAACL-HLT (1)*, 1930–1940.

Harris, Z. S. 1954. Distributional Structure. *WORD* 10(2-3): 146–162.

Hewitt, J.; and Manning, C. D. 2019. A Structural Probe for Finding Syntax in Word Representations. In *NAACL-HLT (1)*, 4129–4138.

Howard, J.; and Ruder, S. 2018. Universal Language Model Fine-tuning for Text Classification. In *ACL (1)*, 328–339.

Huang, L.; Sun, C.; Qiu, X.; and Huang, X. 2019. GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge. In *EMNLP-IJCNLP*, 3507–3512.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *EMNLP (Findings)*, 4163–4174.

Joshi, M.; Chen, D.; Liu, Y.; Weld, D. S.; Zettlemoyer, L.; and Levy, O. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Trans. Assoc. Comput. Linguistics* 8: 64–77.

Kwon, S.; Kang, C.; Han, J.; and Choi, J. 2019. Why Do Masked Neural Language Models Still Need Common Sense Knowledge? *CoRR* abs/1911.03024.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *ICLR*.

Lauscher, A.; Vulic, I.; Ponti, E. M.; Korhonen, A.; and Glavas, G. 2019. Informing Unsupervised Pretraining with External Linguistic Knowledge. *CoRR* abs/1909.02339.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2012. The Winograd Schema Challenge. In *KR*, 552–561.

Levine, Y.; Lenz, B.; Dagan, O.; Ram, O.; Padnos, D.; Sharir, O.; Shalev-Shwartz, S.; Shashua, A.; and Shoham, Y. 2020. SenseBERT: Driving Some Sense into BERT. In *ACL*, 4656–4667.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL*, 7871–7880.

Lin, Y.; Tan, Y. C.; and Frank, R. 2019. Open Sesame: Getting Inside BERT's Linguistic Knowledge. *CoRR* abs/1906.01698.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692.

Loureiro, D.; and Jorge, A. 2019. Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation. In *ACL (1)*, 5682–5691.

Lucy, L.; and Gauthier, J. 2017. Are Distributional Representations Ready for the Real World? Evaluating Word Vectors for Grounded Perceptual Meaning. In *RoboNLP@ACL*, 76–85.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *ACL (1)*, 3428–3448.

McInnes, L.; and Healy, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *CoRR* abs/1802.03426.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, 3111–3119.

Miller, G. A. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38(11): 39–41.

Navigli, R.; and Ponzetto, S. P. 2012. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artif. Intell.* 193: 217–250.

Niven, T.; and Kao, H. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *ACL (1)*, 4658–4664.

Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; and Auli, M. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *NAACL-HLT (Demonstrations)*, 48–53.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, 1532–1543.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *NAACL-HLT (1)*, 2227–2237.

Peters, M. E.; Neumann, M.; Logan, IV, R. L.; Schwartz, R.; Joshi, V.; Singh, S.; and Smith, N. A. 2019. Knowledge Enhanced Contextual Word Representations. In *EMNLP-IJCNLP*, 43–54.

Pilehvar, M. T.; and Camacho-Collados, J. 2019. WiC: The Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *NAACL-HLT (1)*, 1267–1273.

Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H. T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; and Zhong, Z. 2013. Towards Robust Linguistic Analysis using OntoNotes. In *CoNLL*, 143–152.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving Language Understanding with Unsupervised Learning. Technical report, OpenAI.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21: 140:1–140:67.

Reif, E.; Yuan, A.; Wattenberg, M.; Viégas, F. B.; Coenen, A.; Pearce, A.; and Kim, B. 2019. Visualizing and Measuring the Geometry of BERT. In *NeurIPS*, 8592–8600.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *CoRR* abs/1910.01108.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *ICML*, 5926–5936.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*, 4444–4451.

Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; and Wu, H. 2019. ERNIE: Enhanced Representation through Knowledge Integration. *CoRR* abs/1904.09223.

Sun, Z.; Yu, H.; Song, X.; Liu, R.; Yang, Y.; and Zhou, D. 2020. MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices. In *ACL*, 2158–2170.

Tenney, I.; Xia, P.; Chen, B.; Wang, A.; Poliak, A.; McCoy, R. T.; Kim, N.; Durme, B. V.; Bowman, S. R.; Das, D.; and Pavlick, E. 2019. What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. In *ICLR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019a. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *NeurIPS*, 3261–3275.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019b. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *ICLR*.

Wang, C.; Liang, S.; Zhang, Y.; Li, X.; and Gao, T. 2019c. Does it Make Sense? And Why? A Pilot Study for Sense Making and Explanation. In *ACL (1)*, 4020–4026.

Wang, W.; Bi, B.; Yan, M.; Wu, C.; Xia, J.; Bao, Z.; Peng, L.; and Si, L. 2020. StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. In *ICLR*.

Warstadt, A.; Cao, Y.; Grosu, I.; Peng, W.; Blix, H.; Nie, Y.; Alsop, A.; Bordia, S.; Liu, H.; Parrish, A.; Wang, S.; Phang, J.; Mohananey, A.; Htut, P. M.; Jeretic, P.; and Bowman, S. R. 2019. Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. In *EMNLP-IJCNLP*, 2877–2887.

Weaver, W. 1955. Translation. In *Machine Translation of Languages: Fourteen Essays*, 15–23. MIT Press.

Wiedemann, G.; Remus, S.; Chawla, A.; and Biemann, C. 2019. Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *KONVENS*, 161–170.

Wittgenstein, L. 1953. *Philosophical Investigations*. Macmillan.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J. G.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*, 5754–5764.

Ye, Z.; Chen, Q.; Wang, W.; and Ling, Z. 2019. Align, Mask and Select: A Simple Method for Incorporating Commonsense Knowledge into Language Representation Models. *CoRR* abs/1908.06725.

Zellers, R.; Bisk, Y.; Schwartz, R.; and Choi, Y. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *EMNLP*, 93–104.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL (1)*, 4791–4800.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *ACL (1)*, 1441–1451.

Zhou, X.; Zhang, Y.; Cui, L.; and Huang, D. 2020. Evaluating Commonsense in Pre-Trained Language Models. In *AAAI*, 9733–9740.