What Does This Word Mean?
Explaining Contextualized Embeddings with Natural Language Definition

Ting-Yun Chang **Yun-Nung Chen**
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
r06922168@ntu.edu.tw y.v.chen@ieee.org

Abstract

Contextualized word embeddings have boosted many NLP tasks compared with traditional static word embeddings. However, the word with a specific sense may have different contextualized embeddings due to its various contexts. To further investigate what contextualized word embeddings capture, this paper analyzes whether they can indicate the corresponding sense definitions and proposes a general framework that is capable of explaining word meanings given contextualized word embeddings for better interpretation. The experiments show that both ELMo and BERT embeddings can be well interpreted via a readable textual form, and the findings may benefit the research community for a better understanding of what the embeddings capture[1].

1 Introduction

Contextualized word embeddings, such as ELMo, BERT, and OpenAI GPT, GPT-2 (Peters et al., 2018; Devlin et al., 2018; Radford et al., a,b) have been shown to yield richer representations of meaning and boosted many NLP tasks. To understand what contextualized word embeddings capture, Schuster et al. (2019) recently visualized the representations of ELMo and showed that 1) embeddings of the same word in different contexts can form a cluster, and 2) when a word has multiple senses, the embeddings can be separated into multiple distinct groups, one for each meaning. To further investigate the meanings contextualized word embeddings indicate, this paper focuses on analyzing whether a contextualized embedding is sense-informative enough to indicate the corresponding sense definition given a (target word, context) pair. We train and evaluate our model on the online Oxford dictionary dataset released by Chang et al. (2018).

To analyze if the embeddings are sense-informative, our work focuses on learning a mapping between the semantic space of contextualized word embeddings and the space of definition embeddings. Specifically, a better mapping indicates richer sense-specific cues in the contextualized word embedding.

Different from the definition modeling in the prior work (Noraset et al., 2017; Gadetsky et al., 2018; Chang et al., 2018), we reformulate the task from natural language generation (NLG) to classification, i.e., selecting the most reasonable definition according to the target word and its contexts. As recent work has shown the great success in encoding lexical resources into consistent representations (Tissier et al., 2017; Bahdanau et al., 2017; Bosc and Vincent, 2018), in this paper, we leverage pretrained sentence encoder (Cer et al., 2018) to project all definitions in the Oxford dictionary to a consistent embedding space, supporting our reformulation which requires to learn a mapping transforming from the contextualized word embedding space to the definition embedding space. Therefore, we can avoid some predicaments in NLG, such as troubles in generating fluent sequences, the exposure bias problem (Ranzato et al., 2015) and the difficulties in evaluation (Stent et al., 2005).

2 Methodology

The goal of this paper is to analyze whether we can distill sense-specific information from the pretrained contextualized word embeddings such as ELMo and BERT for better interpretation. Specifically, given the embedding of a (word, context) pair, our model learns a non-linear mapping network $f : X \to Y$ to project it into the desired

[1] The source code and the trained models are publicly available at https://github.com/MiuLab/GenDef.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6064–6070,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

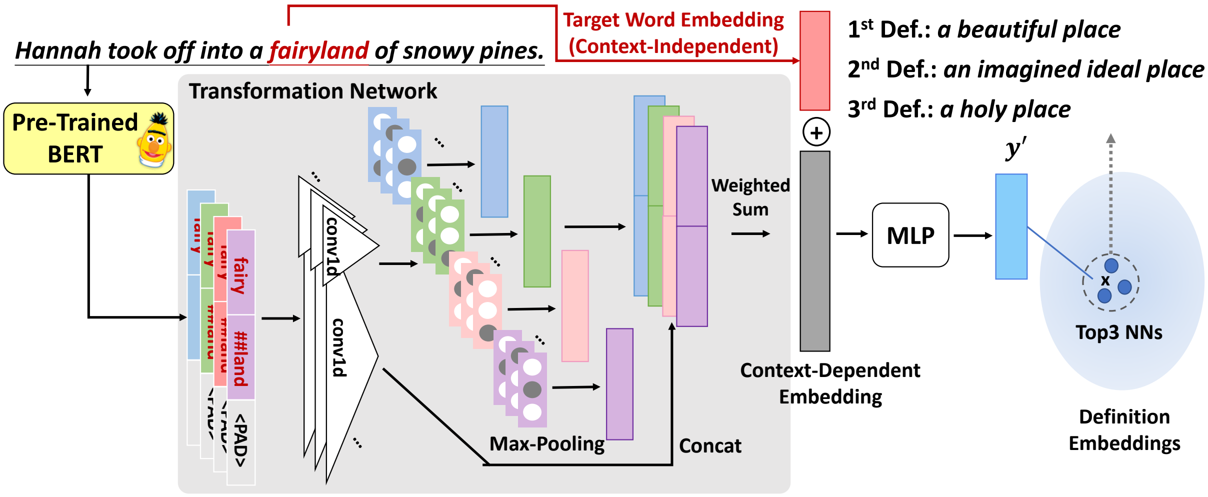

Figure 1: Illustration of the proposed model using contextualized embeddings from BERT as the context-dependent component. We use the one-dimensional convolution with its kernel sizes being 1 and 3.

definition space. Note that the mapping is many-to-one intrinsically because there exist some examples sharing the same definition, and their target words or contexts differ. This paper assumes that *contextualized word embeddings can be easily translated into their corresponding definitions if their semantics is well captured in the representations*. To validate the assumption, the following models are proposed for mapping the embeddings.

2.1 Model Objective

The whole framework of our proposed model is illustrated in Figure 1. Our goal is to learn the mapping that transforms contextualized word embeddings into their corresponding definition, which is to solve:

$$f^\star = \arg\min_f \|f(\mathbf{x}) - \mathbf{y}\|_2, \tag{1}$$

where $\mathbf{x} = (\mathbf{u}, \mathbf{v}) \in X$ consists of the target word embedding $\mathbf{u}$, which is context-independent and serving as the target word identity, and a context-dependent component $\mathbf{v}$, which could be either the context embedding or the contextualized target word embedding; $\mathbf{y} \in Y$ is the embedding of the corresponding definition. Both the definition embedding and the context embedding are encoded by the pretrained transformer-based universal sentence encoder (Cer et al., 2018), and we utilize different types of contextualized word embeddings, such as ELMo, BERT-base, and BERT-large, by swapping the context-dependent component shown in the left part of Figure 1.

Note that though we model this task as a classification problem, we do not train a classifier that regards all definitions as discrete labels but learn a translation between two representation spaces, motivated by Lample et al. (2017, 2018), because different definitions may have semantic similarities. We first encode all ground truth definitions to a consistent embedding space and learn a mapping function f. During the inference stage, given a target word and its context, we retrieve the corresponding human-readable definitions of the top-k nearest neighbors of our predicted embedding in the definition embedding space, with consideration of the whole 79,030 definition candidates in the Oxford dictionary. Note that the candidates for each word are not restricted to its existing definitions in the dictionary, considering that words may go through semantic shift, such as the word *gay* has shifted its meaning from *happy* to *homosexual*.

2.2 Mapping Architecture

Our mapping architecture consists of a transformation network followed by the 7-layer MLP, with batch normalization (Ioffe and Szegedy, 2015) and ReLU. In order to incorporate diverse context-dependent embeddings (such as ELMo and BERT), different transformation nets are proposed, whose common goal is to map the input features to a fixed-dimension representation. Three variants are described in detail.

- **Context Embedding**: the target word embedding and its context embedding encoded by a pretrained sentence encoder are concatenated as the input to the transformation net, which is simply implemented as a fully-connected layer with the ReLU activation.

- **ELMo**: we apply a weighted sum over 3 extracted contextualized word embeddings, i.e., the output of character CNN and two LSTMs, getting a single context-dependent vector for concatenating with the target word vector.

- **BERT**: the target word is tokenized into word pieces, and we use one-dimensional convolution (conv1d) (Kim, 2014) and max-pooling to tackle the variable-lengthed features. We extract the last 4 layers from BERT and jointly learn softmax-normalized weights corresponding to each layer similar to ELMo.[2] Figure 1 illustrates the mapping model leveraging features from BERT, which expresses the best capability of carrying sense-specific explanation among all variants.

2.3 Reverse Mapping

In order to analyze what the mapping captures for better interpretation, we examine the reverse direction of our mapping after training, motivated by Yuan et al. (2016). Given a context-dependent embedding $\mathbf{v}$ and its ground truth definition embedding $\mathbf{y}$, also the word embedding $\mathbf{u}_w$ for each target word w in the vocabulary V, and a pretrained mapping $\bar{f}$, the word that is the closest vector to the target definition after mapping is formulated as:

$$\arg\max_{w \in V} \cos(\bar{f}(\mathbf{u}_w, \mathbf{v}), \mathbf{y}). \qquad (2)$$

In our experiments, the word set corresponding to the top-k highest cosine scores often contains the actual target word, also overlapping with a few synonyms provided by the Oxford dictionary. When applying to contextualized word embeddings from BERT-base, the model achieves the average recall of 23.7%, even though we do not incorporate any synonym information during training. This demonstrates that our models are capable of automatically capturing potential similarities. Examples of the generated synonyms can be found in the supplementary material.

3 Experiments

In order to examine whether the sense-specific information captured by contextualized word embeddings can be well disentangled, the following experiments are conducted.

3.1 Definition Retrieval

This is to analyze whether our proposed mapping indeed interprets the sense-specific definitions from contextualized word embeddings. Considering that the words can be seen and unseen, our experiments contain two levels of tasks (Chang et al., 2018).

- **Seen** is to test the pair with *(seen word, unseen context, seen definition)*, including 151,306 pairs of instances containing 9,276 target words.

- **Unseen** is to test the pair with *(unseen word, unseen context)*, including 15,959 pairs of instances corresponding to the 1,000 randomly selected target words held-out from training. Such a **zero-shot** setting challenges if the input feature is informative enough and if the mapping can generalize to the unseen but semantically consistent embeddings. Also, it is a practical and appealing task as many new words are coined every year.

We ensure both being polysemic tasks by sampling within instances whose target words have at least 3 definitions when building these two test sets.

3.1.1 Results

We measure the performance of various proposed architectures with average precision (@1, @5, @10) as well as the average cosine distance between the predicted definition embedding and the ground truth embedding (lower is better). Two baselines without using contextualized embeddings as context-dependent input features are proposed, 1) using the target word embedding only, which entirely ignores the contexts and thus being a lower bound of this task and 2) leveraging the context embedding from the pretrained Transformer-based universal-sentence encoder as

[2] Note that our work focuses on analyzing the sense information encoded in the contextualized embeddings; thus, our model is stacked upon the frozen representations extracted from the pretrained BERT instead of fine-tuning them. More training details are given in the supplementary material.

Task	Methods	P@1	P@5	P@10	Cosine Dist
Seen	target word embedding	33.27 / 18.29	50.77 / 31.92	56.06 / 36.69	0.251
	+ context embedding	59.36 / 45.19	71.43 / 58.17	74.95 / 62.42	0.178
	+ ELMo	67.00 / 53.91	77.32 / 65.69	80.35 / 69.58	0.149
	+ BERT-base	**74.83 / 63.34**	**83.28 / 73.97**	**85.46 / 77.06**	**0.123**
	+ BERT-large	73.89 / 62.36	82.61 / 73.24	84.92 / 76.28	0.126
Zero-Shot	target word embedding	1.84 / 1.06	6.54 / 4.22	9.67 / 6.44	0.388
	+ context embedding	1.97 / 1.29	7.00 / 4.77	10.78 / 7.50	0.383
	+ ELMo	2.04 / 1.38	6.79 / 4.65	10.21 / 7.06	0.387
	+ BERT-base	3.27 / 2.28	9.59 / 7.41	14.44 / 11.44	0.344
	+ BERT-large	**3.50 / 2.52**	**10.47 / 8.17**	**15.58 / 12.35**	**0.339**

Table 1: Precision@K (average within examples sharing the same target words / average within examples sharing the same (target word, definition)) and cosine distance for models using various input features.

Methods	Tasks	
	Seen	Unseen
Noraset et al. (2017)	21.6 / 36.7	1.7 / 15.8
Chang et al. (2018)	24.9 / 41.0	2.0 / 15.9
target word embedding	28.4 / 36.9	4.6 / 17.2
+ context embedding	58.5 / 62.8	5.1 / 16.8
+ ELMo	66.5 / 71.6	4.8 / 17.2
+ BERT-base	**74.7 / 78.3**	**7.1 / 19.3**

Table 2: BLEU@4 / ROUGE-L:F scores of NLG-based models and various proposed architectures.

described in Section 2.2. Note that the naive baseline is randomly guessing among the whole 79,030 definitions, with P@1 lower than 0.0013%, showing the difficulty of this task.

For **Seen** experiments, Table 1 shows that the context-dependent component contains abundant sense-informative cues, where contextualized word embeddings, especially BERT, expresses the strong capability of producing corresponding definitions with about 15% enhancement of P@1 comparing to the second baseline. For **Unseen** results, the trend is similar: all models with context-dependent input features outperform the first baseline, and the variants with BERT reach the best scores among all metrics. The above results demonstrate rich sense-informative cues captured by the contextualized word embeddings.

Furthermore, we evaluate the definitions by their natural language surfaces using BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) scores. The results are in Table 2, where both Noraset et al. (2017) and the first proposed baseline generates or selects definitions depending merely on the static target word embedding, and all other architectures are context-dependent. We initialize the target word embeddings for all models with pretrained `fasttext` (Joulin et al., 2016) on Wikipedia 2017, UMBC webbase corpus for a fair comparison. The results demonstrate that our mapping model can better explain the word representations than the prior work.

3.1.2 Analysis

An ablation study is conducted in which only the contextualized word embedding is used. While the scores of all 3 related variants drop dramatically due to the lack of the explicit, context-independent signal to the target word identity, the scores still outperform the first baseline by 7% (P@1) on the **Seen** task, showing the superiority of contextualized representations to their static counterparts. In addition, the reason about better performance of the models with contextualized embeddings compared to the one with context embeddings (the second baseline) is that two context embeddings sharing the same target word sense may differ a lot due to various words in the contexts, but they may have similar contextualized word embeddings produced from ELMo or BERT. This allows our proposed model to better interpret the sense information.

Despite the overall low performance under the zero-shot setting, it is found that all proposed models with context-dependent components are still able to disambiguate different senses. Table 3 shows a randomly-sampled example from the output of the BERT-base model. Although the model can only correctly answer the first definition of *draw*, which may be the most common usage so that it can be easily generated from the input embedding, we show that our model is still able to capture the other two very different word senses

Target	Contexts, Selected Definitions, Ground Truth
draw	The embodied capacity to write and *draw* seems to rule over the languid group of objects underneath. **1st Definition**: produce an image of someone or something by making lines and marks on paper **2nd Definition**: produce a picture or diagram by making lines and marks on paper with a pencil pen etc **3rd Definition**: compose or draw up something written or abstract **Ground Truth**: produce an image of someone or something by making lines and marks on paper
	When it came to the end of the day, though, I was more than happy to *draw* the curtains and shut the day out. **1st Definition**: arrange something carefully into a particular shape or position **2nd Definition**: arrange objects or parts in a zigzag formation or so that they are not in line **3rd Definition**: draw a circle round something especially to focus attention on it **Ground Truth**: pull curtains shut or open
	... hinders your ability to impart spin on the ball, reducing your ability to *draw* and fade the shot on command. **1st Definition**: put the ball in play by throwing it up between two opponents **2nd Definition**: strike the ball in the direction of ones followthrough so that it travels to the left ... **3rd Definition**: propel a ball with a bat racket stick etc to score runs or points in a game **Ground Truth**: hit the ball so that it deviates slightly usually as a result of spin

Table 3: The analysis of the top 3 selected definitions on the **Unseen** task.

Model	Accuracy (%)
Lee and Chen (2017)	52.14
Neelakantan et al. (2015)	54.00
Mancini et al. (2016)	54.56
Guo et al. (2019)	55.27
Chang et al. (2018)	57.00
Pilehvar and Collier (2016)	58.55
Proposed (BERT-base)	**68.64**

Table 4: The results on Word-in-Context (WiC) data.

according to the selected definitions. More output samples on both **Seen** and **Unseen** tasks can be found in the supplementary material.

Moreover, unlike the prior work that required discrete token generation to interpret the pretrained word embeddings (Noraset et al., 2017; Gadetsky et al., 2018; Chang et al., 2018), we reformulate the definition modeling task from an NLG problem to a classification problem via learning a mapping between two semantically continuous spaces, which greatly simplifies the hardness, making significant improvement as shown in Table 2. Specifically, as the input representations of Noraset et al. (2017) and the first proposed baseline are the same, i.e., they both are context-agnostic and utilize the same pretrained static word embeddings, the better performance of our model demonstrates the direct benefits of not requiring sequence generation.

3.2 Word Sense Selection in Context

We further examine if the captured sense-specific cues help word sense disambiguation via Word-in-Context data (WiC) (Pilehvar and Camacho-Collados, 2018), in which each instance contains a pair of two contexts sharing a target word, and the task is to decide whether their word senses are the same.[3] To justify that the models are capable of selecting senses encoded in the embeddings, for each pair, our model outputs 10 candidate definitions (top-10 nearest neighbors), and we output TRUE if any definition occurs in both candidate sets, otherwise FALSE. Table 4 shows that the proposed model with contextualized word embeddings outperforms all previous models. We conclude that contextualized word embeddings indeed capture sense-informative cues and our proposed model is capable of interpreting the corresponding senses via definition.

4 Conclusion

This paper proposes a framework that can well interpret the contextualized word embeddings by human-readable sense definitions. The experiments demonstrate that contextualized word embeddings capture the sense-informative cues and the proposed model can better explain the semantics encoded in the representations.

Acknowledgements

We would like to thank Ta-Chung Chi for in-depth discussions and anonymous reviewers for their insightful comments on the paper. This work was financially supported from the Young Scholar Fellowship Program by Ministry of Science and Technology (MOST) in Taiwan, under Grant 108-2636-E-002-003.

[3] To analyze the generalizability, we follow the experimental setting of Guo et al. (2019), where all baselines and the proposed model are pretrained and evaluated directly on the large WiC training set without fine-tuning.

References

Dzmitry Bahdanau, Tom Bosc, Stanisław Jastrzebski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. 2017. Learning to compute word embeddings on the fly. *arXiv preprint arXiv:1706.00286*.

Tom Bosc and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Ting-Yun Chang, Ta-Chung Chi, Shang-Chi Tsai, and Yun-Nung Chen. 2018. xSense: Learning sense-separated sparse representations and textual definitions for explainable word sense networks. *arXiv preprint https://arxiv.org/abs/1809.03348*.

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*.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271.

Fenfei Guo, Mohit Iyyer, Leah Findlater, and Jordan Boyd-Graber. 2019. A differentiable self-disambiguated sense embedding model via scaled gumbel softmax.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Guang-He Lee and Yun-Nung Chen. 2017. MUSE: Modularizing unsupervised sense embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 327–337.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2016. Embedding words and senses together via joint knowledge-enhanced training. *arXiv preprint arXiv:1612.02703*.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient non-parametric estimation of multiple embeddings per word in vector space. *arXiv preprint arXiv:1504.06654*.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of AAAI*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2018. WiC: 10,000 example pairs for evaluating context-sensitive representations. *arXiv preprint arXiv:1808.09121*.

Mohammad Taher Pilehvar and Nigel Collier. 2016. De-conflated semantic representations. *arXiv preprint arXiv:1608.01961*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. a. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. b. Language models are unsupervised multitask learners.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *arXiv preprint arXiv:1902.09492*.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 341–351. Springer.

Julien Tissier, Christopher Gravier, and Amaury Habrard. 2017. Dict2vec: Learning word embeddings using lexical dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*.