2022

TDLR: Top (*Semantic*)-Down (*Syntactic*) Language Representation

Vipula Rawte
University of South Carolina - Columbia

Megha Chakraborty
University of South Carolina - Columbia

Kaushik Roy
University of South Carolina - Columbia

Manas Gaur
University of Maryland Baltimore County

Keyur Faldu
Meta

See next page for additional authors

Publication Info
Published in *36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans*, 2022.

TDLR: Top (*Semantic*)-Down (*Syntactic*) Language Representation

Vipula Rawte
Artificial Intelligence Institute South Carolina
University of South Carolina
vrawte@mailbox.sc.edu

Megha Chakraborty
Artificial Intelligence Institute South Carolina
University of South Carolina
meghac@email.sc.edu

Kaushik Roy
Artificial Intelligence Institute South Carolina
University of South Carolina
kaushikr@email.sc.edu

Manas Gaur
Dept of Computer Science and Engineering
University of Maryland, Baltimore County
manas@umbdc.edu

Keyur Faldu
Meta

Prashant Kikani
Embibe

Hemang Akbari
Embibe

Amit Sheth
Artificial Intelligence Institute South Carolina
University of South Carolina
amit@sc.edu

Abstract

Language understanding involves processing text with both the grammatical and common-sense contexts of the text fragments. The text "I went to the grocery store and brought home a car" requires both the grammatical context (syntactic) and common-sense context (semantic) to capture the oddity in the sentence. Contextualized text representations learned by Language Models (LMs) are expected to capture a variety of syntactic and semantic contexts from large amounts of training data corpora. Recent work such as ERNIE has shown that infusing the knowledge contexts, where they are available in LMs, results in significant performance gains on General Language Understanding (GLUE) benchmark tasks. However, to our knowledge, no knowledge-aware model has attempted to infuse knowledge through top-down *semantics-driven syntactic processing* (Eg: Common-sense to Grammatical) and directly operated on the attention mechanism that LMs leverage to learn the data context. We propose a learning framework *Top-Down Language Representation* (**TDLR**) to infuse common-sense semantics into LMs. In our implementation, we build on BERT for its rich syntactic knowledge and use the knowledge graphs ConceptNet and WordNet to infuse semantic knowledge.

1 Introduction

LMs like BERT [1], RoBERTa [3], T5 [7], GPT2 [6] efficiently learn distributed representations for text fragments such as tokens, entities, and phrases based on statistically likely patterns (syntactic - a text fragment's language context is defined by statistically likely neighbors). The language

Preprint. Under review.

syntax is characterized by grammar rules and the frequency of text fragment co-occurrences reflected in large language corpora. These models outperform human baselines GLUE tasks [10]. LMs implicitly model a broad notion of "common-sense" in large language corpora. This is due to the nature of pattern learning (tending to a "normal" distribution) on large data. However, human understandable semantics found in external knowledge sources such as ConceptNet and WordNet is not explicitly leveraged. We might explicitly leverage the knowledge graph ConceptNet [9] to derive the common-sense conceptual knowledge that world war I and II are different. Distinct concepts would have different neighboring contexts (graphical neighborhoods) in ConceptNet (Eg: world war one-trench warfare, world war two-radio communications). The knowledge graph WordNet [5] gives possible word senses for words. LMs can use the word-sense knowledge from WordNet explicitly to process equivalence between "What does eat the phone battery quickly" and "What would cause the battery on my phone to drain so quickly". The words "eat" and "drain" carry a similar word sense in this example. There has been a growing trend of research around the techniques to infuse knowledge from knowledge graphs into LMs to improve performance [12] [11] [2] [10]. We propose the *Top-Down Language Representation* (**TDLR**) framework - a technique to explicitly infuse common-sense semantics as humans do from available knowledge graphs that capture such semantics. The framework proposes a clear set of steps for *top-down semantics driven syntactic processing* while providing simple mechanisms to expand the scope of the driving semantics utilized. (Eg: Expanding the scope to factual common-sense knowledge such as the current president of a country, found in the knowledge graph WikiData).

2 TDLR Learning Framework

The **TDLR** framework performs three simple steps:

- Construct syntactic representations of the knowledge graphs and the data (Embedding Knowledge and Data at the Syntactic Level).
- Explicitly encode the desired semantics from relevant knowledge graphs in the self-attention mechanism of LMs (Encoding Knowledge Graph Semantics).
- Train the LM as before, thus enabling desired semantics-driven processing of the syntactic information (Knowledge Graph Semantics Driven Syntax Processing).

We show how the **TDLR** framework processes a sentence using the running example: "The World Wars have had a significant impact on 21st-century technology. The great war introduced tanks in battle, and the second world war introduced the use of sophisticated and encrypted radio communications, the drain caused by resource-hungry tech propelled the advancement of modern transistor technology.".

2.1 Embedding Knowledge and Data at the Syntactic Level

The sentence is embedded by deriving and concatenating its constituent word embeddings obtained using a word embedding model [4]. Next, the knowledge concepts are encoded using a knowledge graph embedding technique [8]. Finally, the word embedding and knowledge concept embedding representations are concatenated. For example, the term "War" in our running example has representations from the word2vec (word-embedding model), ConceptNet Numberbatch embedding model, and the convAI WordNet embedding model. Next, all three representations are concatenated to obtain the final representation for the word "war". Finally, all the individual word representations are concatenated to form the sentence representations. Thus we get representations of the sentence that contain the syntactic information from the embedding models.

2.2 Encoding Knowledge Graph Semantics

The word "war" appears in many contexts (Eg: civil war, drug war, proxy war), and the context "world war" may not be so common in the language corpora used to train embedding models. While knowledge graphs like ConceptNet contain the concepts of civil war, drug war, and proxy war in the same graphical context, the embedding models such as Numberbatch have aggregate representations of all the contexts in a given graphical neighborhood, thus losing specific meanings. Therefore we construct a knowledge graph mask that encodes the particular contexts of interest that represent the semantics that will drive the processing of the syntactic input and knowledge representations.

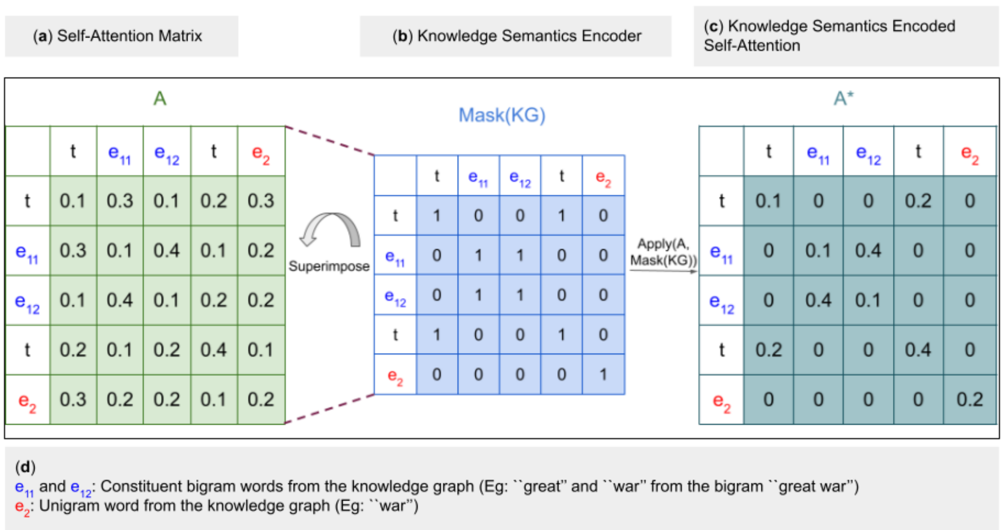

Figure 1: Shows how **TDLR** applied knowledge graph masks to the self-attention mechanism in LMs can explicitly encode graph semantics. Figure 1 (a) shows the self-attention matrix, (b) shows the knowledge graph semantics encoded in a mask, and (c) shows the knowledge encoded self-attention matrix after the mask is applied.

Using our running example, let e_{11} refer to the word "great" and e_{12} refer to the word "war" respectively (see Figure 1 (d)). Assuming that the word "war" has civil, drug, and proxy contexts in the data, an LM trained without explicitly encoding the semantic context "great war" might not capture this meaning. Thus we ensure that the word "war" attends to the word "great" by setting the corresponding entry in the mask to 1 while masking out the rest of the entries with 0 (see Figure 1 (b)). Likewise, denoting the singleton word "war" as e_2 (see Figure 1 (d)), similarly enables knowledge graph semantics to be encoded in the corresponding mask entries for the singleton word "war". In essence, using our approach, we have explicitly encoded the semantic context for the word "war" to mean itself and the accompanying word "great". After encoding the desired semantics in the mask (see Figure 1 (b)), we apply the mask to obtain a knowledge semantics encoded self-attention matrix (see Figure 1 (c)).

Bayesian Perspective: A question might arise that the knowledge semantics encoded self-attention matrix has lost its probabilistic interpretation (the row and column sums are no longer = 1). We can see the application of the mask as a natural application of the Bayes rule in Equation 1.

$$Posterior(A \mid K, data) = \frac{Likelihood(data \mid A)Prior(A \mid K)}{Z} \tag{1}$$

Here A is Self-Attention, K is the knowledge, and Z is the normalizing constant. The posterior in Equation 1 is A^* and the prior is A. The knowledge mask encodes a prior probability distribution (unnormalized as row and column sums are not 1). The self-attention matrix encodes data-likelihood probabilities. Thus we can liken the application of the mask to a likelihood prior product that is proportional to the posterior probability.

2.3 Knowledge Graph Semantics Driven Syntax Processing

With the desired knowledge semantics encoded in the self-attention matrix, we execute the forward-backward training pass as usual in an LM (see Figure 2). Expanding the knowledge semantics scope that drives the top-down processing in **TDLR** requires the simple addition of multiple attention masks at different layers.

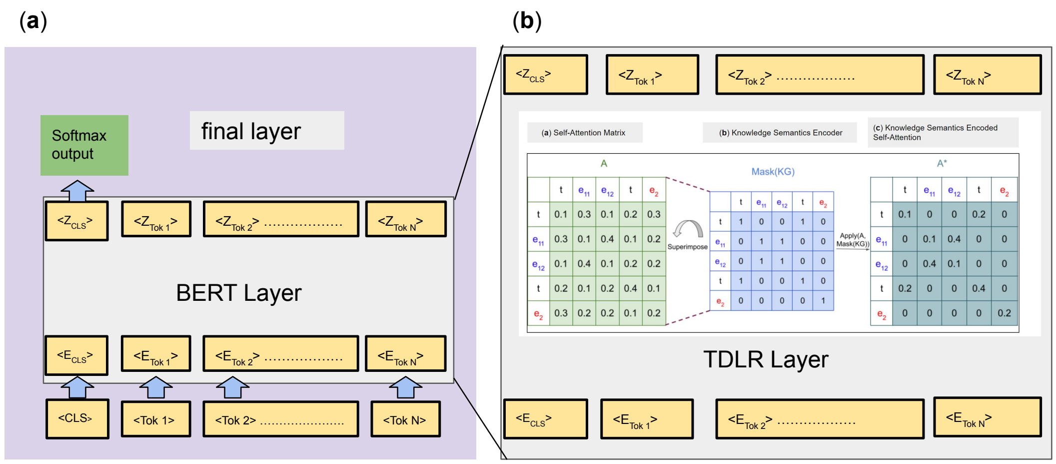

Figure 2: (a) A Transformer LM layer - BERT Layer, and (b) shows the BERT layer with the knowledge semantics encoded self-attention computation.

3 Experiments

We test the **TDLR** method on GLUE benchmark tasks that require the infusion of specific knowledge semantics in the data. We build **TDLR** on the $\text{BERT}_{\text{BASE}}$ model and the $\text{BERT}_{\text{LARGE}}$ model. Both these models execute "normally" distributed semantics driven syntactic processing. To infuse semantics contained in WordNet and ConceptNet, we encode the graph information at the input (syntactic) level (see Section 2.1), as well as apply mask encodings that capture the semantics in these knowledge graphs (see Section 2.2). Thus **TDLR** executes ConceptNet and WordNet semantics-driven processing of the syntactic information in the language for a series of benchmark tasks.

In Table 1 and 2 we see that for tasks that require common-sense semantic knowledge, such as scientific exam questions and identifying conceptual similarities in quora question pairs, even the BASE model of **TDLR** (**TDLR** built on $\text{BERT}_{\text{BASE}}$) outperforms $\text{BERT}_{\text{LARGE}}$. The experiment clearly shows the benefit of targeted re-contextualization achieved through top-level common-sense semantics from WordNet and ConceptNet to drive the processing of the syntactic text inputs. **TDLR** also achieves an average accuracy of **80.46**% across the GLUE Tasks of MNLI, QQP, SST-2, CoLA, STS-B, MRPC, and RTE. Comparatively $\text{BERT}_{\text{LARGE}}$ and $\text{BERT}_{\text{BASE}}$ score **80.17**% and **79.6**% respectively. The GLUE task experiment underscores the performance improvements achieved by using common-sense knowledge for language understanding in general.

Interestingly, varying dataset sizes, as shown in Table 2, also show how **TDLR** needs relatively smaller amounts of data for good performance. Thus, we also see the role of infusing semantics in common-sense knowledge sources to improve performance for low-resource tasks.

System	SciTail	QQP(Academic)	QNLI(Academic)	MNLI(Academic)	Average
$\text{BERT}_{\text{BASE}}$	90.97	71.94	81.64	61.36	76.47
$\text{BERT}_{\text{LARGE}}$	92.89	74.79	84.17	65.15	79.25
$\textbf{TDLR}_{\textbf{BASE}}$	**93.55**	**77.51**	**87.56**	**69.7**	**82.08**

Table 1: Comparing **TDLR** performance on tasks that require common-sense semantic knowledge.

System	Parameters	SciTail(15%)	SciTail(30%)	SciTail(50%)	SciTail(100%)
$\text{BERT}_{\text{BASE}}$	110M	85.74	87.44	90.22	90.97
$\text{BERT}_{\text{LARGE}}$	330M	90.26	91.76	91.25	**92.89**
$\textbf{TDLR}_{\textbf{BASE}}$	111M	**90.82**	**92.28**	**92.05**	**92.89**

Table 2: Comparing **TDLR** performance on different dataset sizes for the SciTail task.

4 Conclusion and future work

We propose Top Down Language Representations (**TLDR**), a method to infuse knowledge in the self-attention mechanism. **TDLR** enables top-level semantics-driven bottom-level language processing at a general level. We demonstrate **TDLR**'s performance improvements using common-sense semantics from WordNet and ConceptNet built on top of BERT. In future work, we will explore extensions that use common-sense semantics, such as factual knowledge in Wikipedia and domain-specific knowledge in the Unified Medical Language System.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang. K-bert: Enabling language representation with knowledge graph. In *AAAI*, 2020.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[5] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995. ISSN 0001-0782. doi: 10.1145/219717.219748. URL `https://doi.org/10.1145/219717.219748`.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

[8] R. Speer and J. Lowry-Duda. Conceptnet at semeval-2017 task 2: Extending word embeddings with multilingual relational knowledge. *arXiv preprint arXiv:1704.03560*, 2017.

[9] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press, 2017.

[10] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://aclanthology.org/W18-5446`.

[11] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf`.

[12] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1139. URL `https://aclanthology.org/P19-1139`.