

2022

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

Vipula Rawte

Megha Chakraborty

Kaushik Roy

Manas Gaur

Keyur Faldu

See next page for additional authors

Follow this and additional works at: https://scholarcommons.sc.edu/aii_fac_pub



Part of the [Computer Engineering Commons](#), and the [Electrical and Computer Engineering Commons](#)

Author(s)

Vipula Rawte, Megha Chakraborty, Kaushik Roy, Manas Gaur, Keyur Faldu, Prashant Kikani, Hemang Akbari, and Amit Sheth

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

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

17 1 Introduction

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

21 syntax is characterized by grammar rules and the frequency of text fragment co-occurrences reflected
22 in large language corpora. These models outperform human baselines GLUE tasks [10]. LMs
23 implicitly model a broad notion of “common-sense” in large language corpora. This is due to
24 the nature of pattern learning (tending to a “normal” distribution) on large data. However, human
25 understandable semantics found in external knowledge sources such as ConceptNet and WordNet is
26 not explicitly leveraged. We might explicitly leverage the knowledge graph ConceptNet [9] to derive
27 the common-sense conceptual knowledge that world war I and II are different. Distinct concepts
28 would have different neighboring contexts (graphical neighborhoods) in ConceptNet (Eg: world
29 war one-trench warfare, world war two-radio communications). The knowledge graph WordNet
30 [5] gives possible word senses for words. LMs can use the word-sense knowledge from WordNet
31 explicitly to process equivalence between “What does eat the phone battery quickly” and “What
32 would cause the battery on my phone to drain so quickly”. The words “eat” and “drain” carry a
33 similar word sense in this example. There has been a growing trend of research around the techniques
34 to infuse knowledge from knowledge graphs into LMs to improve performance [12] [11] [2] [10].
35 We propose the *Top-Down Language Representation (TDLR)* framework - a technique to explicitly
36 infuse common-sense semantics as humans do from available knowledge graphs that capture such
37 semantics. The framework proposes a clear set of steps for *top-down semantics driven syntactic*
38 *processing* while providing simple mechanisms to expand the scope of the driving semantics utilized.
39 (Eg: Expanding the scope to factual common-sense knowledge such as the current president of a
40 country, found in the knowledge graph WikiData).

41 2 TDLR Learning Framework

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

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

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

54 2.1 Embedding Knowledge and Data at the Syntactic Level

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

64 2.2 Encoding Knowledge Graph Semantics

65 The word “war” appears in many contexts (Eg: civil war, drug war, proxy war), and the context
66 “world war” may not be so common in the language corpora used to train embedding models. While

67 knowledge graphs like ConceptNet contain the concepts of civil war, drug war, and proxy war in the
 68 same graphical context, the embedding models such as Numberbatch have aggregate representations
 69 of all the contexts in a given graphical neighborhood, thus losing specific meanings. Therefore we
 70 construct a knowledge graph mask that encodes the particular contexts of interest that represent the
 71 semantics that will drive the processing of the syntactic input and knowledge representations.

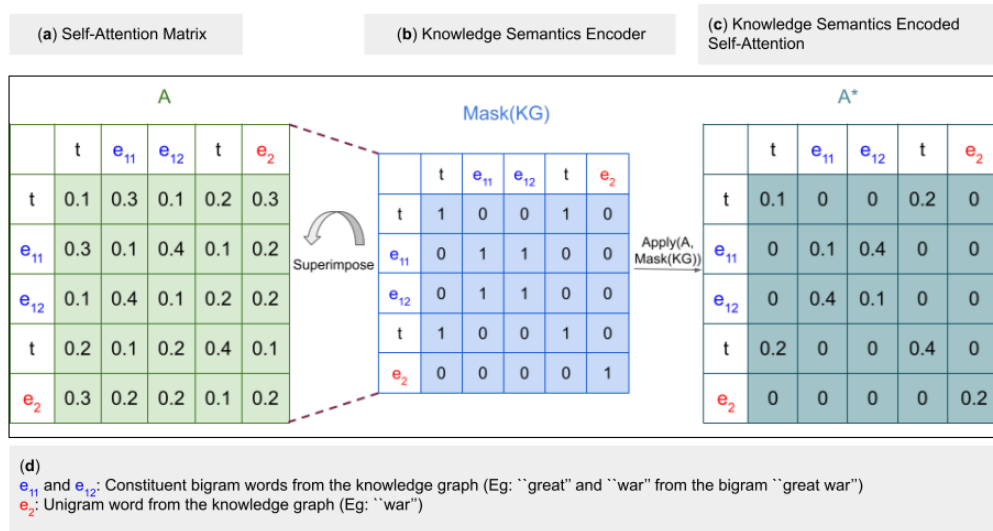


Figure 1: Shows how **TDLR** applied knowledge graph masks to the self-attention mechanism in LMs can explicitly encode graph semantics. Figure 1 (a) shows the self-attention matrix, (b) shows the knowledge graph semantics encoded in a mask, and (c) shows the knowledge encoded self-attention matrix after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

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

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

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

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

91 **2.3 Knowledge Graph Semantics Driven Syntax Processing**

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

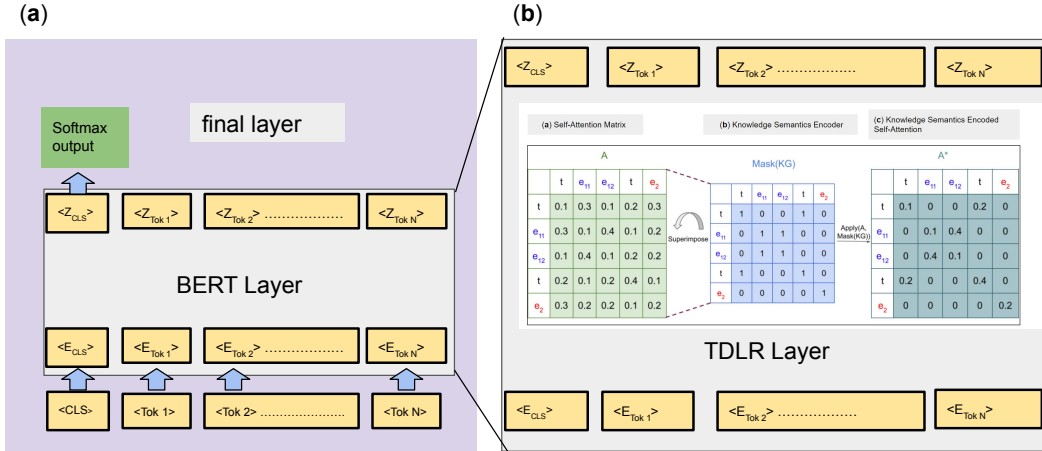


Figure 2: (a) A Transformer LM layer - BERT Layer, and (b) shows the BERT layer with the knowledge semantics encoded self-attention computation.

95

96 **3 Experiments**

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

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

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

System	SciTail	QQP(Academic)	QNLI(Academic)	MNLI(Academic)	Average
BERT _{BASE}	90.97	71.94	81.64	61.36	76.47
BERT _{LARGE}	92.89	74.79	84.17	65.15	79.25
TDLR_{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%)
BERT _{BASE}	110M	85.74	87.44	90.22	90.97
BERT _{LARGE}	330M	90.26	91.76	91.25	92.89
TDLR_{BASE}	111M	90.82	92.28	92.05	92.89

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

117 4 Conclusion and future work

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

124 References

- 125 [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional
 126 transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- 127 [2] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang. K-bert: Enabling language
 128 representation with knowledge graph. In *AAAI*, 2020.
- 129 [3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer,
 130 and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*
 131 *arXiv:1907.11692*, 2019.
- 132 [4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in
 133 vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- 134 [5] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995.
 135 ISSN 0001-0782. doi: 10.1145/219717.219748. URL [https://doi.org/10.1145/219717.](https://doi.org/10.1145/219717.219748)
 136 [219748](https://doi.org/10.1145/219717.219748).
- 137 [6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are
 138 unsupervised multitask learners. 2019.
- 139 [7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J.
 140 Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal*
 141 *of Machine Learning Research*, 21(140):1–67, 2020. URL [http://jmlr.org/papers/v21/](http://jmlr.org/papers/v21/20-074.html)
 142 [20-074.html](http://jmlr.org/papers/v21/20-074.html).
- 143 [8] R. Speer and J. Lowry-Duda. Conceptnet at semeval-2017 task 2: Extending word embeddings
 144 with multilingual relational knowledge. *arXiv preprint arXiv:1704.03560*, 2017.
- 145 [9] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general
 146 knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*,
 147 *AAAI'17*, page 4444–4451. AAAI Press, 2017.

- 148 [10] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task
149 benchmark and analysis platform for natural language understanding. In *Proceedings of the*
150 *2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*,
151 pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi:
152 10.18653/v1/W18-5446. URL <https://aclanthology.org/W18-5446>.
- 153 [11] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and
154 S. Bowman. Superglue: A stickier benchmark for general-purpose language understand-
155 ing systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and
156 R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Cur-
157 ran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper/2019/file/](https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf)
158 [4496bf24afe7fab6f046bf4923da8de6-Paper.pdf](https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf).
- 159 [12] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. ERNIE: Enhanced language rep-
160 resentation with informative entities. In *Proceedings of the 57th Annual Meeting of the*
161 *Association for Computational Linguistics*, pages 1441–1451, Florence, Italy, July 2019.
162 Association for Computational Linguistics. doi: 10.18653/v1/P19-1139. URL [https:](https://aclanthology.org/P19-1139)
163 [//aclanthology.org/P19-1139](https://aclanthology.org/P19-1139).