# Enhancing Question Answering by Injecting Ontological Knowledge through Regularization

**Travis R. Goodwin** and **Dina Demner-Fushman**
U.S. National Library of Medicine
National Institutes of Health
{firstname.lastname}@nih.gov

## Abstract

Deep neural networks have demonstrated high performance on many natural language processing (NLP) tasks that can be answered directly from text, and have struggled to solve NLP tasks requiring external (e.g., world) knowledge. In this paper, we present OSCR (Ontology-based Semantic Composition Regularization), a method for injecting task-agnostic knowledge from an Ontology or knowledge graph into a neural network during pre-training. We evaluated the performance of BERT pre-trained on Wikipedia with and without OSCR by measuring the performance when fine-tuning on two question answering tasks involving world knowledge and causal reasoning and one requiring domain (healthcare) knowledge and obtained 33.3 %, 18.6 %, and 4 % improved accuracy compared to pre-training BERT without OSCR.

## 1 The Problem

"The detective flashed his badge to the police officer." The nearly effortless ease at which we, as humans, can understand this simple statement belies the depth of semantic knowledge needed for its understanding: What is a detective? What is a police officer? What is a badge? What does it mean to *flash* a badge? Why would the detective need to flash his badge to the police officer? Understanding this sentence requires knowing the answer to all these questions and relies on the reader's knowledge about this world: a detective investigates crime, police officers restrict access to the crime scene, and a badge can be a symbol of authority.

As shown in Figure 1, suppose we were interested in determining whether, upon showing the policeman his badge, it is more plausible that the detective would be let into the crime scene or that the police officer would confiscate the detective's badge? To answer this question, we would need to leverage our accumulated expectations about the world: although both scenarios are certainly possible, our accumulated expectations about the world suggest it would be very extraordinary for the police officer to confiscate the detective's badge rather than allow him to enter the crime scene.

**Premise:** The detective flashed his badge to the police officer.

**Question:** What is the most likely *effect*?

**A:** The police officer confiscated the detective's badge.

**B:** The police officer let the detective enter the crime scene.

Figure 1: Example of a question requiring commonsense and causal reasoning (Roemmele et al., 2011) with entities highlighted.

Evidence of Grice's Maxim of Quantity (Grice, 1975), this shared knowledge of the world is rarely explicitly stated in text. Fortunately, some of this knowledge can be extracted from Ontologies and knowledge bases. For example ConceptNet (Speer et al., 2017) indicates that a *detective* is a TypeOf *police officer*, and is CapableOf *finding evidence*; that *evidence* can be LocatedAt a *crime scene*; and that a *badge* is a TypeOf *authority symbol*.

While neural networks have been shown to obtain state-of-the-art performance on many types of question answering and reasoning tasks from raw data (Devlin et al., 2018; Rajpurkar et al., 2016; Manning, 2015), there has been less investigation into how to inject ontological knowledge into deep learning models, with most prior attempts embedding ontological information outside of the network itself (Wang et al., 2017).

In this paper, we present a pre-training regular-



*Proceedings of Deep Learning Inside Out (DeeLIO):*
*The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures,* pages 56–63

ization technique we call OSCR (Ontology-based Semantic Composition Regularization), which is capable of injecting world knowledge and ontological relationships into a deep neural network. We show that incorporating OSCR into BERT's pre-training injects sufficient world knowledge to improve fine-tuned performance in three question answering datasets. The main contributions of this work are:

1. OSCR, a regularization method for injecting ontological information and semantic composition into deep learning models;
2. Empirical evidence showing the impact of OSCR on two tasks requiring world knowledge, causal reasoning, and discourse understanding even with as few as 500 training example, as well as a task requiring medical domain knowledge; and
3. Experimental results showing that the same technique used to infer background knowledge about the world can also capture domain-specific knowledge in the case of medical question answering; and
4. An open-source implementation of OSCR and BERT supporting mixed-precision training, non-TPU model distribution, and enhanced numerical stability.

## 2 Background and Related Work

The idea of training a model on a related problem before training on the problem of interest has been shown to be effective for many natural language processing tasks (Dai and Le, 2015; Peters et al., 2017; Howard and Ruder, 2018). More recent uses of pre-training adapt transfer learning by first training a network on a language modeling task and then fine-tuning (retraining) that model for a supervised problem of interest (Dai and Le, 2015; Howard and Ruder, 2018; Radford et al., 2018). Pre-training, in this way, has the advantage that the model can build on previous parameters to reduce the amount of information it needs to learn for a specific downstream task. Conceptually, the model can be viewed as applying what it has already learned from the language model task when learning the downstream task.

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained neural network that has been shown to obtain state-of-the-art results on eleven natural language processing tasks after fine-tuning (Devlin et al., 2018). BERT relies on two pre-training objectives: (1) a variant of language modeling called *Cloze* (originally proposed in Taylor 1953) where-in 20 % of the words in a sentence are masked, and the model must unmask them and (2) a next sentence prediction task where-in the model is given two pairs of sentences and must decide if the second sentence immediately follows the first. Despite its strong empirical performance, the architecture of BERT is relatively simple: four layers of transformers (Vaswani et al., 2017) are stacked to process each sentence.

In terms of injecting knowledge into pre-training, Zhang et al. (2019) explored injecting entity information into BERT using multi-head attention. However, their approach requires explicitly indicating entity boundaries or relation constituents with special input tokens for down-stream fine-tuning. By contrast, OSCR requires no modification of input formats in the host network. Sun et al. (2019) explored modifying BERT's pre-training by masking entire entities and phrases extracted from external knowledge. Meanwhile, Xie et al. (2019) explored projecting propositional knowledge using Graph Convolutional Networks (GCNs). OSCR, instead, introduces a regularization term that can be added to any natural language pre-training objectives, without modifying the architecture of the network or the pre-training objectives themselves.

## 3 The Data

Incorporating OSCR into pre-training requires an embedded ontology (or knowledge) graph, and one or more natural language pre-training objectives to regularize – in our case, BERT's *Cloze* and next-sentence prediction tasks. These objectives, in turn, require a document collection.

### 3.1 The Ontology

ConceptNet 5 is a semantic network containing relational knowledge contributed to Open Mind Common Sense (Singh et al., 2002) and to DBPedia (Auer et al., 2007), as well as dictionary knowledge from Wiktionary, the Open Multilingual WordNet (Singh et al., 2002; Miller, 1995), the high-level ontology from OpenCyc[1], and knowledge about word associations from "Games with a Purpose" (von Ahn, 2006). In our experiments we used ConceptNet 5 as our ontology relying on an embedded representation of the ontology known as ConceptNet NumberBatch (Speer et al., 2017),

[1] http://www.cyc.com/opencyc/

in which embeddings for all entities in ConceptNet were built using an ensemble of (a) data from ConceptNet, (b) word2vec (Mikolov et al., 2013), (c) GloVe (Pennington et al., 2014), and (d) OpenSubtitles 2016[2] using retrofitting.

### 3.2 The Documents

Our text corpus was a 2019 dump of English Wikipedia articles with templates expanded as provided by Wikipedia's Cirrus search engine[3]. Preprocessing relied on NLTK's Punkt sentence segmenter[4] (Loper and Bird, 2002), and the WordPiece subword tokenizer provided with BERT.

## 4 The Approach

Virtually all neural networks designed for natural language processing represent language as a sequence of words, subwords, or characters. By contrast, Ontologies and knowledge bases encode semantic information about *entities*, which may correspond to individual nouns (e.g., "badge") or multiword phrases ("police officer"). Consequently, injecting world and domain knowledge from a knowledge base into the network requires *semantically decomposing* the information about an entity into the supporting information about its constituent words. For example, injecting the semantics of "Spanish Civil War" into the network requires learning what information the word "Spanish" introduces to the nominal "Civil War" and what information "Civil" adds to the word "War". To do this, OSCR is implemented using a three-step approach illustrated in Figure 2:

**Step 1.** entities are recognized in a sentence using a Finite State Transducer (FST);

**Step 2.** the sequence of subwords corresponding to each entity are semantically composed to produce an entity-level encoding; and

**Step 3.** the average energy between the composed entity encoding and the pre-trained entity encoding from the ontology is used as a regularization term in the pre-training loss function.

By training the model to compose sequences of subwords into entities, during back-propagation, the semantics of each entity are decomposed and injected into the network based on the neural activations associated with its constituent words.

[2] http://opus.nlpl.eu/OpenSubtitles-v2016.php

[3] https://www.mediawiki.org/wiki/Help:CirrusSearch

[4] https://www.nltk.org/_modules/nltk/tokenize/punkt.html

### 4.1 Entity Detection

We designed OSCR to require as few modifications to the underlying host network (e.g., BERT) as possible. We recognized entities during training and inference online by (1) tokenizing each entity in our ontology using the same tokenizer used to prepare the BERT pre-training data, and (2) compiling a Finite State Transducer to detect sequences of subword IDs corresponding to entities. The FST, illustrated in Figure 3, allowed us to detect entities on-the-fly without hard coding a specific ontology and without inducing any discernible change in training or inference time. Although we did not explore it in this work, this potentially allows for multiple ontologies to be injected through OSCR during pre-training. In these experiments, due to the simplicity of ConceptNet entities, we relied on exact string matching to detect entities. Formally, let $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$ represent the sequence of words in a sentence. The FST processes $\boldsymbol{X}$ and returns three sequences: $s_1, s_2, \cdots, s_M$; $l_1, l_2, \cdots, l_M$; and $e_1, e_2, \cdots, e_M$ representing the start offset, length, and the pretrained embedded representation of every mention of any entity in the Ontology.

**Entity Subsumption.** When detecting entities, it is often the case that multiple entities may correspond to the same span of text. As illustrated in Figure 2, the entity "Spanish Civil War" contains the subsumed entities "Spanish", "Civil War", "Civil", and "War". Likewise, because BERT masks 20 % of the words in each sentence, it is possible for entities to involve masked words. Note: including or excluding subsumed and de-masked entities (as illustrated in Figure 2) provided no discernible effect in our experiments.

**Entity Demasking.** Because BERT masks tokens when pre-training, we evaluated the impact of (a) de-masking words before detecting entities and (b) ignoring all entity mentions involving masked words.

### 4.2 Semantic Composition

The role of semantic composition in OSCR, is to learn a composed representation $\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_M$ for each entity detected in $\boldsymbol{X}$ such that $\boldsymbol{c}_i = \mathit{compose}\left(x_{s_i}, x_{s_i+1}, \cdots, x_{s_i+l_i}\right)$. As pre-training in

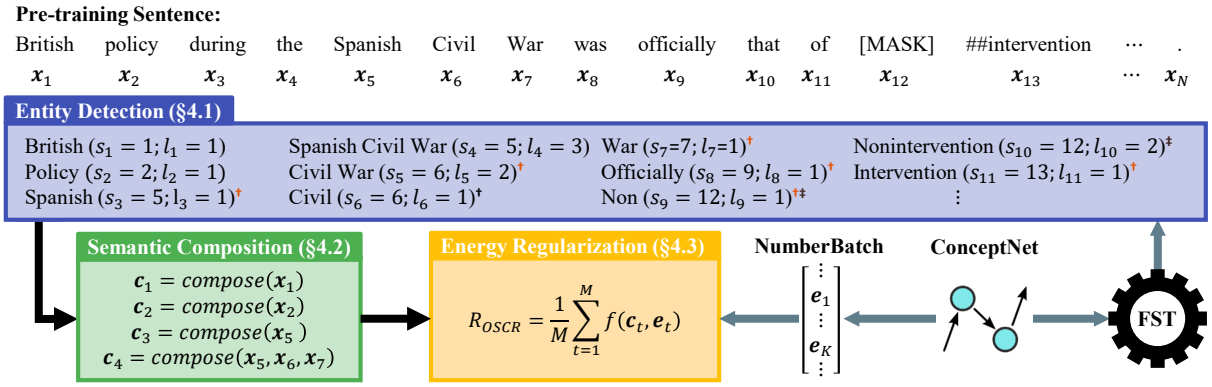


Figure 2: Architecture of OSCR when injecting ontology knowledge from ConceptNet into BERT where '†' indicates subsumed entities, '‡' indicates de-masked entities, $N$ is the length of the input sentence, $M$ is the number of entities detected in the sentence, and $K$ is the number of entities with embeddings in ConceptNet.

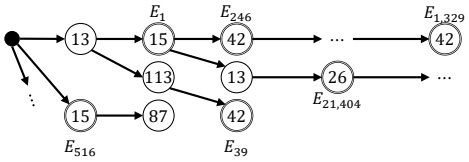


Figure 3: Finite State Transducer (FST) used to detect entities during pretraining; each node corresponds to a word ID, double circles represent terminal states, and $e_i$ indicates the $i^{\text{th}}$ pretrained entity embedding in ConceptNet's NumberBatch.

BERT is computationally expensive, we considered three computationally-efficient methods for composing words and subwords into entities.

**Recurrent Additive Networks** (RANs) are a simplified alternative to LSTM- or GRU-based recurrent neural networks that use only additive connections between successive layers and have been shown to obtain similar performance with 38% fewer learnable parameters (Lee et al., 2017).

Given a sequence of words $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_L$ we use the following layers to accumulate information about how the semantics of each word in an entity contribute to the overall semantics of the entity:

$$\widetilde{\boldsymbol{m}}_t = \boldsymbol{W}_m \boldsymbol{x}_t \tag{1a}$$
$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_i \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_i\right) \tag{1b}$$
$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_f \left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_f\right) \tag{1c}$$
$$\boldsymbol{m}_t = \boldsymbol{i}_t \circ \widetilde{\boldsymbol{m}}_t + \boldsymbol{f}_t \circ \boldsymbol{m}_{t-1} \tag{1d}$$
$$\boldsymbol{h}_t = g\left(\boldsymbol{m}_t\right) \tag{1e}$$

where $[\bullet]$ represents vector concatenation, $\widetilde{\boldsymbol{m}}_t$ represents the content layer which encodes any new semantic information provided by word $\boldsymbol{x}_t$, $\circ$ indicates an element-wise product, $\boldsymbol{i}_t$ represents the input gate, $\boldsymbol{f}_t$ represents the forget gate, $\boldsymbol{m}_t$ represents the internal memories about the entity, and $\boldsymbol{h}_t$ is the output layer encoding accumulated semantics about word $\boldsymbol{x}_t$. We define the composed entity $\boldsymbol{c}_i \coloneqq \boldsymbol{h}_{s_i+l_i}$ (i.e., the content vector of the RAN after processing the last token in the entity) for the sequence beginning with $\boldsymbol{x}_{s_i}$.

**Linear Recurrent Additive Networks** To further reduce model complexity, we considered a second, simpler version of a RAN omits the content and output layers (i.e., Equations 1a and 1e) and Equation 1d is updated to depend on $\boldsymbol{x}_t$ directly: $\boldsymbol{m}_t = \boldsymbol{i}_t \circ \boldsymbol{x}_t + \boldsymbol{f}_t \circ \boldsymbol{m}_{t-1}$. As above, we define the composed entity $\boldsymbol{c}_i \coloneqq \boldsymbol{m}_{s_i+l_i}$ for the sequence of subwords beginning with $\boldsymbol{x}_{s_i}$.

**Linear Interpolation** Finally, we considered a third, even simpler form of semantic composition. Inspired by Goodwin and Harabagiu (2016), we represented the semantics of an entity as an unordered linear combination of the semantics of its constituent words, i.e.: $\boldsymbol{c}_i \coloneqq \boldsymbol{W}_e\left(\boldsymbol{x}_{s_i} + \boldsymbol{x}_{s_i+1} + \cdots + \boldsymbol{x}_{s_i+l_i}\right) + l_i \cdot \boldsymbol{b}_e$.

### 4.3 Energy Regularization

We project the composed entities into the same vector space as the pretrained entity embeddings from the Ontology, and measure the average energy across all entities detected in the sentence:

$$\mathcal{R}_{\text{OSCR}} = \frac{1}{M}\sum_{i=1}^{M} f\left(\boldsymbol{W}_p \boldsymbol{c}_i + \boldsymbol{b}_p, \boldsymbol{e}_i\right) \tag{2}$$

where $f$ is an *energy function* capturing the energy between the composed entity $c_i$ and the pretrained entity embedding $e_i$. We considered three energy

functions: (1) the Euclidean distance, (2) the absolute distance, and (3) the angular distance, which can handle negative values.

## 5 Experiments

### 5.1 Experimental Setup

**Hyper-parameter Tuning** For each fine-tuning task, we used a greedy approach to hyper-parameter tuning by incrementally and independently optimizing: batch size $\in \{8, 16, 32\}$; initial learning rate $\in \{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}\}$; whether to include subsumed entities $\in$ {yes, no}; and whether to include masked entities $\in$ {yes, no}.

For CoPA, the Story Cloze task, and RQE, we found an optimal batch size of 16 and an optimal learning rate of $2 \times 10^{-5}$. We also found that including subsumed entities and masked was optimal (at a net performance improvement of $< 1\%$ accuracy).

**Pretraining** We pretrained BERT using a 2019 Wikipedia dump formatted for Wikipedia's Cirrus search engine.[5] Preprocessing relied on NLTK's Punkt sentence segmenter[6] (Loper and Bird, 2002), and the WordPiece subword tokenizer provided with BERT. We used the vocabulary from BERT base (not large) and a maximum sequence size of 384 subwords, training 64 000 steps, with an initial learning rate of $2 \times 10^{-5}$, and 320 warm-up steps.

**BERT Modifications** We used a modified version of BERT, allowing for mixed-precision training. This necessitated a number of minor changes to improve numerical stability around softmax operations. Training was performed using a single node with 4 Tesla P100s each (multiple variants of OSCAR were trained simultaneously using five such nodes at a time). Non-TPU multi-GPU support was added to BERT based on Horovod[7] and relying on Open MPI.

### 5.2 Results

We evaluated the impact of OSCR on three question answering tasks requiring world or domain knowledge and causal reasoning.

**Choice of Plausible Alternatives** a SemEval 2012 shared task, (CoPA) presents 500 training and 500 testing sets of two-choice questions and requires to choose the most plausible cause or effect entailed by the premise, as illustrated in Figure 1 (Roemmele et al., 2011). The topics of these questions were drawn from two sources: (1) personal stories taken from a collection of blogs (Gordon and Swanson, 2009); and (2) subject terms from the Library of Congress Thesaurus for Graphic Materials, while the incorrect alternatives were created so as to penalize "purely associative methods".

**Premise:** Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.

**Ending A:** She found her phone in the car.

**Ending B:** She didn't want her phone anymore.

Figure 4: Example of a Story Cloze question (correct answer is A).

**Consumer Health Question:** Can sepsis be prevented. Can someone get this from a hospital?

**FAQ A:** Who gets sepsis?

**FAQ B:** What is the economic cost of sepsis?

Figure 5: Example of a Recognizing Question Entailment (RQE) question (correct answer is A).

**The Story Cloze Test** evaluates story understanding, story generation, and script learning and requires a system to choose the correct ending to a four-sentence story, as illustrated in Figure 4 (Mostafazadeh et al., 2016). In our experiments, we used only the 3,744 labeled stories.

**Recognizing Question Entailment** Healthcare questions can be highly complex compared to general open-domain questions, potentially involving accounting for family, social, and medical history. A proposed solution to healthcare question complexity is to decompose the question into simpler subquestions, which can be more easily answered. Recognizing Question Entailment (RQE, Ben Abacha and Demner-Fushman 2016) consists of 8588 training and 302 testing pairs of consumer health questions (CHQs) and frequently asked questions (FAQs) with labels indicating whether answering the FAQ

[5] https://www.mediawiki.org/wiki/Help:CirrusSearch

[6] https://www.nltk.org/_modules/nltk/tokenize/punkt.html

[7] https://eng.uber.com/horovod/

| Model | CoPA | Cloze | RQE |
|---|---|---|---|
| Cirrus BERT | 55.2 | 74.200 | 74.834 |
| Cirrus BERT + OSCR | **73.6** | **87.974** | **77.815** |
| Composition: RAN | 60.6 | 85.890 | **77.815** |
| Composition: Attention | **73.6** | **87.974** | 75.497 |
| Composition: Linear | 72.8 | 85.516 | 76.490 |
| Energy: Absolute | **72.0** | 83.431 | 75.497 |
| Energy: Euclidean | 60.6 | 85.890 | 75.497 |
| Energy: Angular | 59.2 | **86.264** | **77.815** |

Table 1: Accuracy of fine-tuned BERT after pretraining on the Cirrus Wikipedia data with and without OSCR.

entails answering the CHQ, as illustrated in Figure 5.

Table 1 presents the results of BERT when pretrained on Wikipedia with and without OSCR, the state-of-the-art, and the average performance of different semantic composition methods and energy functions when calculating OSCR.

## 6 Discussion

### 6.1 The Impact of External Knowledge.

It is clear from Table 1 that incorporating OSCR provided a significant improvement in accuracy for both common sense causal reasoning tasks, indicating that OSCR was able to inject useful world knowledge into the network. We also evaluated the impact of OSCR on the Stanford Question Answering Dataset (SQuAD), version 1.1, and observed no discernable change in performance (an Accuracy of 86.6 % without and 86.5 % with OSCR). The lack of impact of SQuAD is unsurprising, as the vast majority of SQuAD questions can be answered directly by surface-level information in the text, but it shows that injecting world knowledge with OSCR does not come at the expense of model performance for tasks that require little outside knowledge.

### 6.2 The Impact of Domain Knowledge.

While less pronounced than the general domain, for the clinical domain, OSCR provided a modest improvement over standard BERT, and both improved over the state-of-the-art.

### 6.3 The Impact of Entity Masking

**Entity Subsumption** We evaluated the impact of including subsumed entities when calculating OSCR and found it provided, on average, only a minor increase in accuracy (< 1 % average relative improvement) at a 10 % increase in total training time. Consequently, we recommend ignoring all subsumed entities.

**Entity De-masking** De-masking entities had little over-all impact on model performance (< 1% average relative improvement) and no discernible effect on training time. This may be explained by the fact that Wikipedia sentences are typically much longer than standard English sentences, so the likelihood of an important entity being masked is relatively small.

### 6.4 The Role of Semantic Composition

When comparing semantic composition methods, the Linear method had the most consistent performance across both domains; the Recurrent Additive Network (RAN) obtained the lowest performance on the general domain and the highest performance on medical texts, while the Linear RAN exhibited the opposite behavior. While this suggests more complex domains require more complex representations of semantic composition, we recommend Linear composition as it exhibits consistent performance and requires 50% less training time than the RAN and 40% less than the Linear RAN.

### 6.5 The Impact of the Energy Functions

In terms of energy functions, the Euclidean distance was the most consistent, the Angular distance was the best for the Story Cloze and RQE tasks, and the Absolute difference was the best for CoPA. The Angular distance (being scale-invariant) is least affected by the number of subwords constituting an entity while the Absolute distance is most affected. Consequently, we believe the Absolute distance was only effective on the CoPA evaluation because the entities in CoPA are typically very short (single words or subwords). We recommend selecting the energy function based on the average length of entities in the fine-tuning tasks: Angular distance with long entities, Absolute distance with short entities, and Euclidean distance with varied entities.

Finally, we compared the impact of including and excluding subsumed and masked entities and found that neither resulted in any substantial change in model improvements (< 1 % change in accuracy), while ignored masked and subsumed entities lead to a 20 % average reduction in training time.

### 6.6 Limitations and Future Work

In this study, we only considered ConceptNet as our ontology because we were primarily interested in in-

jecting common-sense world knowledge. However, OSCR is not specific to any Ontology. Likewise, we considered only one type of pretrained entity embeddings: ConceptNet NumberBatch (Speer et al., 2017), despite the availability of other, more sophisticated approaches for knowledge graph embedding including, TransE (Bordes et al., 2013), TranR (Lin et al., 2015), TransH (Wang et al., 2014), RESCAL (Nickel et al., 2011) and OSRL(Xiong et al., 2018). In future work, we hope to explore the impact of incorporating different Ontologies and knowledge graphs as well as alternative types of entity embeddings (Bordes et al., 2013; Lin et al., 2015; Wang et al., 2014; Nickel et al., 2011; Xiong et al., 2018).

## 7 Conclusions

In this paper we presented OSCR (Ontology-based Semantic Composition Regularization), a learned regularization method for injecting task-agnostic knowledge from an Ontology or knowledge graph into a neural network during pretraining. We evaluated the impact of including OSCR when pretraining BERT with Wikipedia articles by measuring the performance when fine-tuning on two question answering tasks involving world knowledge and causal reasoning and one requiring domain (healthcare) knowledge and obtained 33.3 %, 18.6 %, and 4 % improved accuracy compared to pre-training BERT without OSCR.

## Reproducibility

All code, data, and experiments are available on GitHub at `https://github.com/h4ste/oscar`.

## Acknowledgments

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health, and utilized the computational resources of the NIH HPC Biowulf cluster (`http://hpc.nih.gov`).

## References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg. Springer-Verlag.

Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2787–2795. Curran Associates, Inc.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3079–3087. Curran Associates, Inc.

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*.

Travis Goodwin and Sanda Harabagiu. 2016. Embedding open-domain common-sense knowledge from text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4621–4628, Portorož, Slovenia. European Language Resources Association (ELRA).

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*, volume 46.

H Paul Grice. 1975. Logic and conversation. *1975*, pages 41–58.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics.

Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2017. Recurrent additive networks. *arXiv preprint arXiv:1705.07393*.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2181–2187. AAAI Press.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher D Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849. Association for Computational Linguistics.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 809–816, USA. Omnipress.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, pages 1223–1237, Berlin, Heidelberg. Springer-Verlag.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration.

Ole Tange. 2018. *GNU Parallel 2018*. Ole Tange.

Wilson L. Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

L. von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Q. Wang, Z. Mao, B. Wang, and L. Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pages 1112–1119. AAAI Press.

Yaqi Xie, Ziwei Xu, Mohan S Kankanhalli, Kuldeep S Meel, and Harold Soh. 2019. Embedding symbolic knowledge into deep networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4233–4243. Curran Associates, Inc.

Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2018. One-shot relational learning for knowledge graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1980–1990. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. 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. Association for Computational Linguistics.