What does KnowBert-UMLS forget?

Guilhem Piat, Nasredine Semmar, Julien Tourille, Alexandre Allauzen, Hassane Essafi

What does KnowBert-UMLS forget?

Guilhem Piat*†, Nasredine Semmar*, Julien Tourille*, Alexandre Allauzen†, Hassane Essafi*
* Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
† Université Paris Dauphine, F-75775, Paris Cedex 16, France

***Abstract*—Integrating a source of structured prior knowledge, such as a knowledge graph, into transformer-based language models is an increasingly popular method for increasing data efficiency and adapting them to a target domain. However, most methods for integrating structured knowledge into language models require additional training in order to adapt the model to the non-textual modality. This process typically leads to some amount of catastrophic forgetting on the general domain. KnowBert is one such knowledge integration method which can incorporate information from a variety of knowledge graphs to enhance the capabilities of transformer-based language models such as BERT. We conduct a qualitative analysis of the results of KnowBert-UMLS, a biomedically specialized KnowBert model, on a variety of linguistic tasks. Our results reveal that its increased understanding of biomedical concepts comes at the cost, specifically, of *general common-sense knowledge* and *understanding of casual speech*.**

***Index Terms*—Domain Adaptation, Knowledge based systems, Catastrophic Forgetting, Machine learning, Biomedical informatics**

I. Introduction

The studies of catastrophic forgetting suffered by knowledge-enhanced language models (Piat *et al.* [1], Xu *et al.* [2], and others) have thus far focused on quantitative analyses of performance, which provide little insight into how to improve future models. This analysis attempts to *qualitatively* characterize the scope of what the models forget, in order to definitively identify these models' weaknesses in hopes of mending them in the future.

Most knowledge graph integration methods require the Knowledge Base (KB) concepts being mentioned in the input text to be identified ahead of time. This assumes the task of Entity Linking (EL), *i.e.* identifying named entities and finding the corresponding KB concept, is solved. This is an unverified assumption, as the state of the art in biomedical Entity Linking is Bhowmik *et al.*'s dual encoder [3] which achieves 0.564 F_1 score on the MedMentions [4] corpus.

KnowBert [5] is a BERT-based language model which alleviates this issue by including the EL task in its training objective. In this paper, we therefore focus our qualitative analysis specifically on KnowBert-UMLS (Piat *et al.* [1]), a language model that incorporates the Unified Medical Language System (UMLS) biomedical knowledge base using the KnowBert method.

We introduce the idea of Knowledge Integration and review the literature on the topic in section II. Section III is dedicated to a more detailed description of the KnowBert approach and KnowBert-UMLS model. We discuss qualitative aspects of the difference in performance between KnowBert-UMLS and baselines in section IV. Lastly, we draw our conclusions and discuss implications for future research in section V.

II. Related Work

Transformer-based language models, when trained mostly on general text such as is the case with BERT [6] and GPT [7], for which the bulk of the training corpus is formed by the *Books* corpus [8], do not perform well on tasks involving specific domains such as medicine, patents, or law. This is to be expected, as many of the writing conventions differ by field and make assumptions on the knowledge of the reader. The obvious way to expand the capabilities of a language model to a specific domain is therefore to include in-domain text in the model's training corpus such as with GeoBERT [9], LEGAL-BERT [10], or PatentBERT [11], as well as BioBERT [12], BlueBERT [13], and ClinicalBERT [14] in the biomedical domain alone.

While this method generally performs well [15], it does have several drawbacks. First, language models are inefficient at learning factual information. This is evidenced by the fact that they are not reliably able to predict facts that appear in their training data [16]. Consequently, models specialized by pre-training on in-domain text typically require several billion words of in-domain text, as with the approximately 21 billion words in the training corpus for BioBERT and 57 billion words for LEGAL-BERT.

Additionally, Arumae and Bhatia [17] and Xu *et al.* [2] have shown that extending pre-training on in-domain text leads to catastrophic forgetting on general language. This can be reduced by balancing in-domain with general language in the specialization corpus, at the cost of more computation and diluting the specialized text, which in turn tends to decrease performance on in-domain tasks.

Another class of approaches to the problem of domain adaptation involves incorporating knowledge from a KB. Typically, this involves identifying domain-relevant entities mentioned in the model's input text, matching them to the concepts recorded in a Knowledge Graph (KG), extracting the relevant knowledge, and enhancing the contextualized word representations with this knowledge. This is the general approach followed by models such as ERNIE [18], KnowBert [5], E-BERT [19], K-Adapter [20], KEPLER [21], UmlsBERT [22], DRAGON [23], and CODER [24].

In theory, this type of approach does not inherently require re-training the model. These approaches therefore tend to be more data-efficient than in-domain pre-training, and tend to cause less catastrophic forgetting as shown by Piat *et al.* [1].

However, the knowledge extracted from the graph typically takes the form of concept embeddings, or the graph may be used to define a knowledge-driven training objective. In all practical cases, some form of training is therefore required for the incorporation of knowledge into contextualized word embeddings to be successful and some catastrophic forgetting is inevitable.

III. KnowBert-UMLS

KnowBert is a knowledge-integration approach for language models which avoids the problem of requiring an entity linker by taking as input, rather than specific entity matches, a set of *candidate spans* which mark *possible* entity mentions in the text, and for each of the candidate spans, a set of candidate concept embeddings from the KB. For instance, given the following sentence:

> *Type 1 diabetes is an autoimmune disease that primarily affects pancreatic internal secretion.*

It may be fed the following candidate spans (outlined in boxes, with subscripts signifying the number of candidate concepts for the given span):

[*Type*]$_1$ *1* [*diabetes*]$_2$]$_1$ *is an*
[*autoimmune*]$_1$ [*disorder*]$_2$]$_1$ *that primarily*
[*affects*]$_2$ [*pancreatic*]$_3$ [*internal*]$_1$ [*secretion*]$_1$]$_1$.

Each of these candidate spans s would be paired with a set of candidate entities $\mathcal{E}_s$. Each entity is defined by its 7-digit *Concept Unique Identifier* and a list of accepted names (*e.g.* "DNA" and "Deoxyribonucleic Acid" would be two names for one concept), one of which is deemed "preferred". Following is an example of a possible set of candidate entities for the candidate span *pancreatic* in the form *Concept Unique Identifier* : *preferred concept name*.

- C0030274: Pancreas
- C0030292: Pancreatic Hormones
- C0030304: Pancreatin

Each of these candidate concepts would be associated with a concept embedding.

Having overlapping and nested candidate spans with multiple candidate concepts per span is a form of entity linking which increases recall at the cost of precision. This ensures that as much relevant information as possible is given to the language model. In order to mitigate the effects of low precision, KnowBert introduces the idea of a soft entity linking by using attention to learn to estimate the posterior probability for each candidate concept of being actually mentioned given the current context. This information is then used as a basis for the knowledge integration.

KnowBert-UMLS is based on the KnowBert architecture, but specifically uses UMLS as its knowledge base. In reality, due to the inaccuracy of the candidate concept identification step, KnowBert-UMLS does not use one concept embedding per concept in UMLS, of which there are approximately 4.6 million. Rather, the concepts are grouped by *Semantic Type*, of which there are 135. This drastically increases precision and recall of the candidate generation step at the cost of decreasing the granularity of information that can be integrated.

TABLE I
Pre-Training Corpus Size (Billions of Words) by Type for Baselines Versus KnowBert-UMLS.

Model	Biomedical	General	Knowledge integration
$\text{BERT}_{\text{BASE}}$	0.0	3.1	No
BioBERT	18.0	3.1	No
PubMedBERT	3.2	0.0	No
BlueBERT	4.5	3.1	No
UmlsBERT	18.5	3.1	Yes
KnowBert-UMLS	2.2	3.1	Yes

IV. Experiments

Following the quantitative results of Piat *et al.* [1], we study the types of mistakes made by KnowBert-UMLS on three tasks: within the biomedical domain, we consider the n2c2 2010 (previously known as i2b2 2010) [25] Named Entity Recognition (NER) task[1]. For our out-of-domain tasks, we study linguistic acceptability with the CoLA task [26] and natural language inference with the SNLI task [27]. As the labels of the test set are not made public for the CoLA task, in order to perform the qualitative analysis, we follow the alternate dataset split introduced by Piat *et al.* [1] (see section IV-C).

To contextualize this qualitative analysis, we recall quantitative results for each task from [1], with multiple baselines which vary in amount of in- and out-of-domain pre-training text and on whether or not they integrate a structured knowledge base. A breakdown of these aspects of these baselines is provided in Table I. For each task, scores are rescaled from $[0, 1]$ to $[0, 100]$ for readability. In tables II, III, and V, the bold score is the highest performance on the task by a specialized model. BERT, as expected, consistently performs highest on general tasks and lowest on n2c2.

As the quantitative results are averages over multiple experiments, a comprehensive analysis of predictions would not be possible. In order to qualitatively analyze the predictions of these models, we fine-tune our pre-trained models to the task once and collect the final predictions on the test set. The specific fine-tuned model used for this analysis may therefore over- or under-perform relative to the reported average scores.

A. Biomedical NER

In this task, models must identify locations where named entities are mentioned and tag them as *problem*s, *test*s or *treatment*s. As recorded in Table II, despite the candidate spans and candidate entities providing it with additional information on named entities in the text, KnowBert-UMLS has the lowest recall among the models we assess. This is likely due to

[1]We exclude the ChemProt task from our analysis as the weaknesses of KnowBert-UMLS reflect lack of biomedical knowledge (as opposed to catastrophic forgetting) which lies outside the scope of this paper.

TABLE II
PERFORMANCE ON THE N2C2 2010 NER TASK.

Model	P	R	F_1
$BERT_{BASE}$	82.71	86.21	84.42
BioBERT	85.20	87.74	86.46
PubMedBERT	86.62	88.28	87.44
BlueBERT	86.68	88.71	87.68
UmlsBERT	86.92	89.46	**88.18**
KnowBert-UMLS	86.63	85.84	86.23

TABLE III
PERFORMANCE ON THE SNLI TASK, MICRO-F_1.

Model	micro F_1
$BERT_{BASE}$	89.24
BioBERT	88.90
PubMedBERT	88.81
BlueBERT	88.20
UmlsBERT	88.59
KnowBert-UMLS	**89.03**

the discrepancy between the terms chosen by the candidate generator and the model for enrichment, and the parts of speech expected to be labeled for this task. Specifically, the n2c2 corpus requires possessive pronouns, determiners, articles, adjectives and other qualifiers to be included in the entity, whereas UMLS (and therefore the candidate generator) requires the opposite. For instance, the following sentences occur in the n2c2 dataset:

1) On postop day number two she was also afebrile$_{\mathbf{1}}$ and had not passed any flatus$_{\mathbf{2}}$ yet.
2) Administer iron products a minimum of 2 hours before or after a levofloxacin$_{\mathbf{3}}$ or ciprofloxacin dose$_{\mathbf{4}}$ [...]

The named entities which are expected to be marked are underlined and numbered with subscripts. The second and third entity mentions include determiners (*any* and *a* respectively) which manual model output examination reveals are not identified by KnowBert-UMLS, whereas the main words in the entity mentions (*flatus* and *levofloxacin* respectively) are accurately identified and tagged. In addition, the candidate generator identifies the main words as entities, but not *a* nor *any*. This is by far the predominant type of false negative within the reviewed set of samples, indicating that the model may have low confidence on terms that do not benefit from knowledge integration.

In comparison, BlueBERT's most prominent weaknesses seem to be that it misses entity mentions entirely or doesn't include all of the words which the mention comprises. In the aforementioned example, for instance, BlueBERT does not recognize the second mention (*any flatus*) as an entity, and it excludes the word 'dose' in the fourth mention. BlueBERT also struggles with determiners, though to a lesser degree. In particular, we have not been able to find an example where BlueBERT misses a possessive pronoun.

Lastly, while this is not a highly prominent issue for either model, KnowBert-UMLS confuses entity types more often than BlueBERT, with approximately 12.7% of KnowBert-UMLS' mistakes being of this kind, versus 8.9% for BlueBERT.

B. General NLI

In the SNLI task, two sequences, a *premise* and *hypothesis*, are fed to the LM. It must determine whether the premise entails the hypothesis, whether they are contradictory, or whether the relationship between sequences is neutral. Surprisingly given BioBERT's performance on the similar WNLI task according to Arumae and Bhatia [17], catastrophic forgetting does not seem to cause substantial degradation in performance, as all specialized models perform fairly well. This is likely due to SNLI not being as adversarial as the WNLI task. KnowBert-UMLS, however, performs best among biomedical models, which is consistent with a reduction of catastrophic forgetting leading to better performance.

KnowBert-UMLS and BERT perform with a high degree of similarity, quantitatively as well as qualitatively, as the specific instances of KnowBert-UMLS and BERT used in this analysis share the same prediction on 94.53% of the instances in the test data. Both models slightly underperform compared to their respective averages, with micro-F_1 scores of 87.99 and 88.33 respectively. However, many of the examples in the SNLI dataset are open to interpretation. For instance:

> Premise: *A bearded man wearing a blue shirt and white t-shirt is working on a fishing net.*
> Hypothesis: *Someone is preparing to catch fish.*

While the hypothesis is the most likely explanation for the premise, it is not difficult to imagine scenarios where the premise is true and the hypothesis is false. In order to capture this ambiguity, in addition to the reference label, each sentence has been manually labeled by five human reviewers. In this case, the label is 'entailment', but two out of five reviewers labeled the relationship between these sentences as 'neutral'. If we accept the reviewers' answers as valid predictions, the micro-F_1 scores of KnowBert-UMLS and BERT (that is, the specific instances used in this analysis) on this task are 96.80 and 96.88 respectively, meaning only approximately one in four errors made by KnowBert-UMLS and BERT were in disagreement with all reviewers.

We decide to examine the mistakes made by both models and attempt to identify patterns. No clear tendencies could be found regarding the incorrect label predictions, but analyzing the instances themselves, we could group mistakes into four major types:

- *Blunders*, for which no information other than what is stated in the text is required to make a decision.
- *Common Sense* (CS) mistakes, which require some reasoning and/or a non-trivial piece of real-world knowledge.
- *Technically Correct* (TC) predictions, in which the model's answer could be considered correct based on an arguably valid interpretation of the text.
- *Not-an-Error* (NaE), for which we agree with the model and disagree with the label, or the input text contains a

TABLE IV
BREAKDOWN OF MISTAKES MADE BY KNOWBERT-UMLS AND BERT BY TYPE ON THE SNLI TASK.

Model	Blunders	CS	TC	NaE
$\text{BERT}_{\text{BASE}}$	38.9%	33.7%	12.9%	14.9%
KnowBert-UMLS	38.6%	38.1%	7.3%	15.7%

major corruption.

We provide examples of the aforementioned categories in Appendix A. After manual examination of the models' failure cases, we report a breakdown of mistakes by type in Table IV.

The main recurring pattern seems to be for KnowBert-UMLS to lack real-world knowledge but stick to more straightforward interpretations of sentences. Illustrative examples for CS and TC mistakes are given in appendices B and C respectively. This may indicate that KnowBert-UMLS suffers from some amount of catastrophic forgetting on real-world knowledge, and perhaps is less prone to noticing and fixating on details which could skew its understanding of the text.

C. Linguistic Acceptability

Our dataset for the Linguistic Acceptability task is based on the CoLA task from the GLUE benchmark. Since CoLA does not make the labels of its test split public however, qualitative analysis of the results could typically not be performed. Following Piat *et al.* [1], we use the validation split for final testing, and replace the validation split with the final 500 entries of the train split as provided in version 1.1 of the dataset.

The objective for this task is to classify sequences as "linguistically acceptable" (*i.e.* correct and written as a native speaker would have) or not. Because F_1 does not account for true negatives in binary classification, models are evaluated using Matthew's Correlation Coefficient (MCC) as is standard with this task.

As shown in Table V, KnowBert-UMLS far outperforms specialized baselines on this task, demonstrating the effectiveness of this method in avoiding catastrophic forgetting. Out of the 527 samples in our test set, the specific instance of KnowBert-UMLS used in this analysis yielded 64 false positives and 25 false negatives for an MCC of 58.35. In comparison, our BERT instance suffered 64 false positives and 20 false negatives for an MCC of 60.89.

Inspection of false negative instances reveals that some of the labels in the corpus reflect types of phrasing that are uncommon in modern written English. For instance:

Came right in he did without so much as a knock.

This phrasing is highly irregular outside of some areas of the UK and lacks punctuation.

Will he can do it?

This sentence uses double modals, which is a nonstandard construction seldom appearing outside of oral speech in Scotland, Northern Ireland and Northern England.

TABLE V
PERFORMANCE OF BERT-BASED LANGUAGE MODELS ON THE MODIFIED COLA TASK, MEASURED AS MATTHEWS' CORRELATION COEFFICIENT (MCC).

Model	MCC
$\text{BERT}_{\text{BASE}}$	60.50
BioBERT	49.30
PubMedBERT	42.90
BlueBERT	39.76
UmlsBERT	44.24
KnowBert-UMLS	**58.52**

Rusty talked about himself only after Mary did talk about him.

While grammatically correct, the use of *did talk* rather than the straightforward simple past *talked* is uncommon in this context. This is likely another regional variance.

While it could be argued that general models should be able to handle non-standard constructions for applications such as speech-to-text transcription or applications involving transcribed text, it can be desirable for a specialized model such as KnowBert-UMLS to classify them as incorrect as such sentence structures may be particularly unlikely to be intentional constructions by a native speaker in the contexts in which the model is susceptible to be deployed.

We therefore manually re-label these false negatives as true negatives in order to estimate the performance of KnowBert-UMLS in these types of contexts. We do not consider re-labeling false positives or true negatives as there are fewer examples of this occurring in the negative class and the practical interpretation would be unclear. Examination of the true positives did not yield any instances which we believe would have warranted relabeling. The full list of incorrect predictions for BERT and KnowBert-UMLS can be found in appendix D, with re-labeled instances marked by a right-facing arrow ('$\rightarrow$').

Taking these new labels into account, the MCC of Knowbert-UMLS' predictions on this task increases to 64.70, outperforming BERT's average performance. While the improved performance demonstrated by this interpretation of the data is meaningful, this score is not necessarily representative of real-world performance, as it neglects several factors. First, performing this relabeling on the train and validation sets in addition to the test set would likely give us a more accurate performance estimate. Second, this is only done for one trained model; results should be averaged over multiple experiments to be representative.

Furthermore, this comparison is (by design) unfair as we place different expectations on both models. When we relabel the false negatives predicted by BERT, its score increases by 2.35 fewer points than KnowBert-UMLS, to 64.89 MCC. KnowBert is therefore not only more prone to rejecting sentences as we can tell from its greater number of negative predictions, but is specifically more prone to rejecting non-standard constructions, which may be a desirable feature.

V. Conclusions

A. Results

The shortcomings of KnowBert-UMLS in the biomedical domain seem to reflect some lack of precision in biomedical knowledge, but more significantly, a difficulty grasping the tasks' expectations regarding which parts of speech to include within named entities. This is in line with the expectation that KnowBert's increased knowledge must come at the cost of some level of proficiency with grammar, and may reflect some level of heterogeneity and inaccuracy of information introduced by KnowBert's knowledge integration process. For out-of-domain tasks, KnowBert-UMLS seems to suffer from a lack of common-sense knowledge, but its slightly degraded performance on CoLA may reflect a desirable compromise which benefits adaptation to biomedical language.

B. Future work

In light of these observations, perhaps the most significant weakness of Knowbert-UMLS is a lack of common-sense knowledge. As Peters *et al.* have shown, a feature of the KnowBert architecture is that it can accommodate multiple knowledge graphs. Adding support for a general knowledge graph such as Wikipedia or a common-sense knowledge graph such as CSKG [28] or ATOMIC [29] could improve the performance of KnowBert-UMLS, particularly on general language and out-of-domain tasks.

Another way of increasing performance in the biomedical domain, particularly to increase the exactitude of the entity types identified, may be to find a better compromise between the high granularity (*i.e.* informativity) of UMLS concepts and the higher accuracy of the candidates and availability of training data on Semantic Types, perhaps by clustering related concepts or falling back onto semantic types only for the concepts on which the candidate generator is highly uncertain. Knowbert-UMLS may also be further specialized in the biomedical domain with the integration of additional KBs such as OpenTargets' LINK.

Beyond the biomedical domain, KnowBert may even support multi-specialization, using knowledgebases from multiple fields such as YAGO [30] or WorldKG [31]. If this is shown to be possible, this method may prove to be a computationally affordable way to increase the breadth of knowledge of large language models such as GPT-3 [32], which could then serve the needs of a greater variety of professionals from different fields.

Lastly, the improvements brought by this method of knowledge integration may be orthogonal to the improvements brought by extended pre-training or other knowledge integration methods. In fact, improved representations of biomedical text may help the knowledge integration module to reach the full extent of its capabilities. A comparative study of performance on a variety of in- and out-of-domain tasks by multiple KnowBert-UMLS-like models with different pretrained LM backbones such as RoBERTa, BioBERT, BlueBERT, and UmlsBERT would shed a valuable light on this topic and may lead to a new state of the art in biomedical language modeling.

Acknowledgment

This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-De-France Regional Council.

References

[1] G. Piat, N. Semmar, A. Allauzen, H. Essafi, G. Bernard, and J. Tourille, "Adapting without forgetting: KnowBert-UMLS," in *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2022, pp. 19–24.

[2] Y. Xu, X. Zhong, A. J. J. Yepes, and J. H. Lau, "Forget me not: Reducing catastrophic forgetting for domain adaptation in reading comprehension," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[3] R. Bhowmik, K. Stratos, and G. de Melo, "Fast and effective biomedical entity linking using a dual encoder," in *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, 2021, pp. 28–37.

[4] S. Mohan and D. Li, "MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts," *arXiv:1902.09476 [cs]*, Feb. 2019.

[5] M. E. Peters, M. Neumann, R. L. Logan, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith, "Knowledge enhanced contextual word representations," in *EMNLP*, 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[7] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," *OpenAI blog*, 2018. [Online]. Available: https://openai.com/research/language-unsupervised

[8] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[9] Y. Gao, Y. Xiong, S. Wang, and H. Wang, "Geobert: Pre-training geospatial representation learning on point-of-interest," *Applied Sciences*, vol. 12, no. 24, p. 12942, 2022.

[10] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of law school," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Associa-

tion for Computational Linguistics, Nov. 2020, pp. 2898–2904.

[11] J.-S. Lee and J. Hsiang, "Patent classification by fine-tuning BERT language model," *World Patent Information*, vol. 61, p. 101965, 2020.

[12] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Sep. 2019.

[13] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 58–65.

[14] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly available clinical bert embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 72–78.

[15] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Jul. 2020, pp. 8342–8360.

[16] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, "Language models as knowledge bases?" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Nov. 2019, pp. 2463–2473.

[17] K. Arumae and P. Bhatia, "CALM: Continuous Adaptive Learning for Language Modeling," *arXiv preprint arXiv:2004.03794*, 2020.

[18] 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*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1441–1451.

[19] N. Poerner, U. Waltinger, and H. Schütze, "E-BERT: Efficient-yet-effective entity embeddings for bert," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2020, pp. 803–818.

[20] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, J. ji, G. Cao, D. Jiang, and M. Zhou, "K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters," *arXiv:2002.01808 [cs]*, Dec. 2020.

[21] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, "KEPLER: A unified model for knowledge embedding and pre-trained language representation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 176–194, 2021.

[22] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, and A. Wong, "UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Jun. 2021, pp. 1744–1753.

[23] M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. S. Liang, and J. Leskovec, "Deep bidirectional language-knowledge graph pretraining," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 309–37 323, 2022.

[24] Z. Yuan, Z. Zhao, H. Sun, J. Li, F. Wang, and S. Yu, "CODER: Knowledge-infused cross-lingual medical term embedding for term normalization," *Journal of biomedical informatics*, vol. 126, p. 103983, 2022.

[25] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.

[26] A. Warstadt, A. Singh, and S. R. Bowman, "Neural network acceptability judgments," *arXiv preprint arXiv:1805.12471*, 2018.

[27] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

[28] F. Ilievski, P. Szekely, and B. Zhang, "CSKG: The commonsense knowledge graph," in *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*. Springer, 2021, pp. 680–696.

[29] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "ATOMIC: An atlas of machine commonsense for *If-Then* reasoning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3027–3035.

[30] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge," in *16th International Conference on the World Wide Web*, 2007, pp. 697–706.

[31] A. Dsouza, N. Tempelmeier, R. Yu, S. Gottschalk, and E. Demidova, "WorldKG: A world-scale geographic knowledge graph," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 4475–4484.

[32] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

APPENDICES

A. Examples of error types for the SNLI task

Pre. A crowd of people looking up at 3 people on the edge of the roof of a building.
Hyp. The crowd on the ground is watching 3 people on the roof's edge.
True: entailment — Predicted: contradiction
This is a Blunder*: the hypothesis is unambiguously a paraphrase of the premise, yet the model predicts* contradiction.

Pre. A group of young men in a gym take turns scoring in basketball.
Hyp. Guys are playing shirts vs skins.
True: neutral — Predicted: contradiction
This is a Common Sense *mistake. KnowBert-UMLS' conception of* shirts vs skins *seems to contradict* basketball. *In the absence of the knowledge of this team differentiation scheme's applicability to most sports, this is a reasonable assumption.*

Pre. Male in a blue jacket decides to lay in the grass.
Hyp. The guy wearing a blue jacket is laying on the green grass.
True: entailment — Predicted: neutral
This is a case of a Technically Correct *answer: the hypothesis is quite clearly a rephrasing of the premise, but adds the detail that the grass is* green *despite it not being strictly necessarily the case.*

Pre. A snowboarder on a wide plain of snow.
Hyp. A snowboarder gliding over a field of snow.
True: neutral — Predicted: entailment
We consider this to be Not an Error. *While arguments can be made for the sentences not being strictly equivalent, we find that it is more reasonable to consider them as such rather than not.*

In the following example, a word which (presumably) contains a typo is underlined, and we subsequently specify what we expect to have been the intended word in brackets.

Pre. A football layer *[player]* wearing a red shirt.
Hyp. A built man wearing a tshirt.
True: neutral — Predicted: contradiction

This is likely due to the wordpiece tokenization scheme and lack of context. *'Layer'* and *'player'* are considered single tokens ; as there is no overlap between the word pieces, the model does not have any information on the similar spellings of the words and cannot be expected to recognize, much less correct, the mistake. We consider this *NaE*.

B. KnowBert-UMLS failure cases on SNLI involving lack of common-sense knowledge

Pre. A girl playing soccer in a green field with some trees in the background.
Hyp. The soccer ball is chasing the girl.
True: contradiction — Predicted: neutral
KnowBert-UMLS does not seem to grasp that the ball chasing the player is not typically a part of the game of soccer.

Pre. An old shoemaker in his factory.
Hyp. The shoemaker is getting ready for his 16^{th} birthday.
True: contradiction — Predicted: neutral
KnowBert-UMLS does not pick up on the contradiction between "old" and "16^{th} birthday".

Pre. Many children play in the water.
Hyp. The children are playing mini golf.
True: contradiction — Predicted: neutral
KnowBert-UMLS does know that mini golf is not typically played in water and predicts neutral.

BERT also fails on all of these examples but the prediction is not always the same as KnowBert-UMLS.

C. Technically correct answers by BERT on the SNLI task

Pre. A boy dressed in a plaid kilt with a brown hat wields a long pole.
Hyp. The boy is holding a samurai sword.
True: contradiction — Predicted: neutral
One could wield both a long pole and a samurai sword, but this is an unlikely scenario.

Pre. A roofer in a gray sweatshirt and orange hat walks on a unfinished roof at a lake-side home.
Hyp. The roofer is putting on shingles.
True: neutral — Predicted: contradiction
Shingling and walking are mutually exclusive actions at any given time, but the action in the hypothesis should be interpreted as a continuous process.

KnowBert-UMLS makes correct predictions for these instances.

D. BERT and KnowBert-UMLS predictions on the CoLA task

This section lists all of the false positive and false negative predictions by BERT and KnowBert-UMLS on the CoLA task. These predictions are separated into six itemized lists depending on whether they are false positive or false negative predictions, and whether they were made by BERT, KnowBert-UMLS, or both.

False negatives which use '→' as a bullet, despite being technically correctly labeled when considering all regional variations of spoken English, are considered mislabeled and are re-labeled as *true* negatives for the purpose of estimating model performance in a literary setting (See section IV-C).

a) Shared false positives:

- The more you would want, the less you would eat.
- The more does Bill smoke, the more Susan hates him.
- Mickey looked up it.
- The tube was escaped by gas.
- What the water did to the bottle was fill it.
- What the water did to the whole bottle was fill it.
- Mary beautifully plays the violin.
- Mary intended John to go abroad.

- Which report that John was incompetent did he submit?
- The mayor regarded as being absurd the proposal to build a sidewalk from Dartmouth to Smith.
- I want that Bill left to remain a secret.
- Drowning cats, which is against the law, are hard to rescue.
- The proof this set is recursive is difficult.
- I live at the place where Route 150 crosses the Hudson River and my dad lives at it too.
- Which hat did Mike quip that she never wore?
- I won't have some money.
- Here's a knife with which for you to cut up the onions.
- The younger woman might have been tall and, and the older one definitely was, blond.
- That the cops spoke to the janitor about it yesterday is terrible, that robbery.
- No writer, and no playwright, meets in Vienna.
- No writer, nor any playwright, meets in Vienna.
- No one can forgive that comment to you.
- This flyer and that flyer differ apart.
- The jeweller scribbled the contract with his name.
- Cynthia chewed.

b) BERT false positives:

- As you eat the most, you want the least.
- I demand that the more John eat, the more he pays.
- We wanted to invite someone, but we couldn't decide who to.
- This is the book which Bob reviewed, and this is the one which Fred won't do it.
- The madrigals which Henry plays the lute and sings sound lousy.
- I can't remember the name of somebody who had misgivings.
- Paperback books lift onto the table easily.
- The books lifted onto the table.
- The chair pushed.
- Did Calvin his homework?
- If I am a rich man, I'd buy a diamond ring.
- The kennel which Mary made and Fido sleeps has been stolen.
- Mary wonders that Bill will come.
- What did you ask who saw?
- Which king did you ask which city invaded?
- Anson became a muscle bound.

c) KnowBert-UMLS false positives:

- Who does John visit Sally because he likes?
- The box contained the ball from the tree.
- Sue gave to Bill a book.
- I know which book José didn't read for class, and which book Lilly did it for him.
- The farmer dumped the cart with apples.
- Herman whipped the sugar and the cream.
- My heart is pounding me.
- I squeaked the door.
- The fort fluttered with many flags.
- John is easy to please Kim.
- Fed knows which politician her to vote for.
- John heard that they criticized themselves.
- Medea tried the nurse to poison her children.
- How fierce the battle?
- The monkey is ate the banana
- I would like to could swim

d) Shared false negatives:

→ The tank leaked the fluid free.

- I know which book Mag read, and which book Bob said that you hadn't.
- We elected me.

→ Sally is tall, and may be blond, and Sheila is short, and definitely is, blond.

- We investigated the area for bombs.
- We recommend to eat less cake and pastry.

→ John bought a book on the table.

- I read some of the book.
- It isn't because Sue said anything bad about me that I'm angry.

→ The man who Mary loves and Sally hates computed my tax.

→ Came right in he did without so much as a knock.

- I saw even the student.

→ Will he can do it?

- I shaved myself.

e) BERT false negatives:

- Jessica loaded boxes on the wagon.
- Carla slid the book.
- Susan whispered at Rachel.
- It is a golden hair.

→ John promise Mary to shave himself.

→ I might be not going to the party but washing my hair

f) KnowBert-UMLS false negatives:

- The mechanical doll wriggled itself loose.

→ Clearly, John probably will immediately learn French perfectly.

→ Rusty talked about himself only after Mary did talk about him.

- I won't ask you to believe that he tried to force me to give her any money.
- The gardener grew that acorn into an oak tree.
- After reading the pamphlet, Judy threw it into the garbage can.
- The boy in the doorway waved to his father.
- That dog is so ferocious, it even tried to bite itself.
- Ann may spend her vacation in Italy.

→ She asked was Alison coming to the party.

→ It is some disgruntled old pigs in those ditches that humans love to eat.