Internet-NLP: Allowing Natural Language Processing Models to Connect to the Internet

Thamognya Kodi
Ruamrudee International School
contact@thamognya.com

Abstract

In this paper, I present **Internet-NLP** a new control-flow wrapper abstraction to enable the utilization of data from the internet (or a knowledge-database when offline) for existing context-needing Natural Lnaguage Processing (NLP) models to function without any given context. Internet-NLP can be used, finetuned alongside existing NLP models via its config settings and additionally its Long Short Term Memory neural network (LSTM neural network) can also be trained. Additionally incorporations of Masked Language Models (MLM) such as BERT, or LinkBERT (Devlin et al., 2019; Yasunaga et al., 2022a) can be utilized to improve search queries, and therfore retrieve more accurate and reliable data. Futhermore, **Internet-NLP** utilizes a LSTM, Reinforcement Learning and caches to allow for multi-turn NLP tasks, and improvement via Reinforcement Learning from user.

Additionally in this paper, I also present new NLP and Natural Language Inference (NLI) models to assist **Internet-NLP**:

- Open-book question and long answer (QA) via GPT-NeoX-20b (Andonian et al., 2021; Black et al., 2022)
- CrossEncoder NLI via LinkBERT (Reimers and Gurevych, 2019; Thakur et al., 2021; Yasunaga et al., 2022a)
- Answer to context NLP via T5 (Raffel et al., 2019)

Along with these models, I also present new general purpose QA and NLI datasets:

- ALotNLI made from ANLI, MultiNLI, and SNLI (Nie et al., 2020; Williams et al., 2018; Bowman et al., 2015)
- ALotOpenBookQA made from CoQA, Natural Questions, and SQuAD (Reddy et al., 2018; Kwiatkowski et al., 2019; Rajpurkar et al., 2018)

As a result of these models, datasets, and Internet-NLP, the accuracy and reliability of most context-needing NLP models on most NLP tasks, especially tasks that require more factual responses with no given context increased.

Internet-NLP and the new NLP and NLI models, which were trained on the general-purpose datasets (ALotNLI, and ALotOpenBookQA). Internet-NLP, by default utilizes an Text-Generative model GPT-NeoX (Andonian et al., 2021; Black et al., 2022) for long responses and LinkBERT (Yasunaga et al., 2022a) for short responses. For 2 choices (for ex: True and False) Bi-Encoder NLI has been used and for multiple choices CrossEncoder will be used (Thakur et al., 2021).

Internet-NLP, in layperson terms, provides the context for context-needing NLP models to let them function. Internet-NLP can be improved via finetuning, and training of LSTM and Reinforcement Learning model (which can be trained alongside the NLP model), which enables for better search queries, and subsequently results. It obtains state-of-the-art (SOTA) results in QA and NLI without context.

Internet-NLP is a subset of a larger package, Internet-ML and is open-source. [1]. Old versions of Internet-NLP is also publicly available. [2].

1 Introduction

There are currently two main solutions for utilizing NLP tasks with no context provided:

1. Large Pre-Trained Text-Generation and Text2Text-Generation Model
    - Pre-trained Text-generation models, like GPT-NeoX, GPT-3, and etc. (Black et al., 2022; Andonian et al., 2021; Brown et al., 2020) can be

[1] Internet-NLP, subset of Internet-ML is public, and open-source: https://github.com/thamognya/internet_ml

[2] Old Versions of Internet-NLP is public: https://pypi.org/project/internet-nlp/

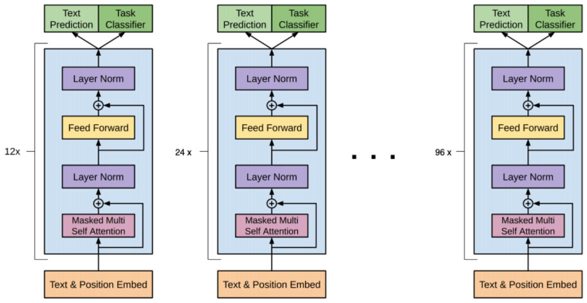

Figure 1: This is an illustration of the architecture of GPT-2 and GPT-3, a popular Text-Generation model (Dzlab; the).

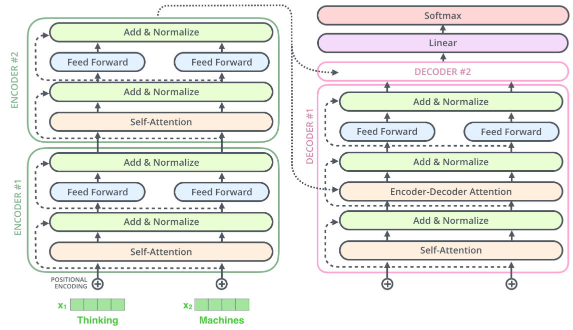

Figure 2: This is an illustration of the architecture of T5, a popular Text2Text-Generation model (Alammar).

trained for open-domain question-answering closed-book language model tasks (ODQA LM) (Weng, 2020). When used in ODQA tasks, they achieve SOTA results in such tasks, have high accuracy and are fast but are much larger in size than open-book (context-needing) language models.

- Additionally Pre-trained Text2Text-generation models, like T5 (Raffel et al., 2019) that have open-domain question-answering closed-book (no context) language models (ODQA LM) capabilities (Weng, 2020). These closed-book QDQA LMs are comparatively state-of-the-art performance in many no-context NLP tasks, mainly question-answering. Text2Text-generation models for such no-context NLP tasks are usually large, slow, and have a low accuracy (Roberts et al., 2020).
- Example: T5 (Raffel et al., 2019)
- Illustration of how ODQA LM work: 1

2. Large Knowledge Database with a Contex-Needing Language Model

- Large Knowledge base with an pre-trained open-book LM and retriever, provides an comparatively higher performance, accuracy and the model itself is small. These models however require — usually — a large knowledge base which makes the overall solution large, but still fast and with more accuracy on the field the knowledge base specalizes in.

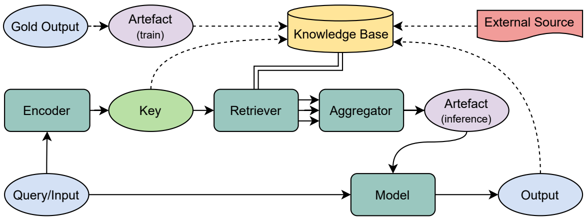

Figure 3: This is an illustration of how LMs with a knowledge base and artifact retriver work (Zouhar et al., 2022).

- Example: LinkBERT (Yasunaga et al., 2022b) with an artifact retriver (Zouhar et al., 2022) with a knowledge base such as DBpedia or WikiData (Auer et al., 2007; Vrandečić and Krötzsch, 2014)
- Illustration of how LMs with a knowledge base and artifact retriver work: 3

Solution 1 and 2 achieve the same end goal of NLP tasks without context via two different methods; these current solutions restrict NLP tasks and accuracy without context, especially for more open-domain tasks. The major limitation in this case would be accuracy, efficency and size of models and their knowledge base which then limit the use cases of closed-book open-domain NLP tasks.

In this paper, I propose Internet-NLP, an direct improvement to solution 1 which allows NLP models to not require a large knowledge base (altough you can configure Internet-NLP to utilize a knowledge base) that incoproates the internet's vast knowledge along utilizing data in hyperlinks in webpages (Yasunaga et al., 2022b) to create a more resource-filled data for our existing or future context-needing pre-trained model to use for NLP tasks. Internet-NLP encompasses pre-trained NLP and NLI models, along with its web-data-scraper creates an small temporary on-basis data and a cache for NLP tasks to be performed without given context.

Utilizing the vast data on the internet, graph of documents as corpus (Yasunaga et al., 2022b) allows us to enable to reduce our solution size, increase efficency and increase accuracy. Additionally unlike usage of static data, Internet-NLP utilizes the dynamic, and frequent updating data of the internet which enables us to utilize any type of NLP model along with NLI models to allow us to follow a sequence of control flow to get the context for context-needing models. This approach utilizes a combination of data-collection (Hrkút et al., 2020) for NLPs with context-needing open-domain NLP to gain more accurate results in most no-context NLP tasks.

Additionally Internet-NLP's Text2Text-generation search query model: T5 (Raffel et al., 2019) and LSTM noun remembrance using parts of speech tagging (Chiche and Yitagesu, 2022) on ALotClosedBookQA with it improving search queries based on the difference on answer recieved and the answer from datasets, using parts of speech tagging on answers (Chiche and Yitagesu, 2022).

References

How deep is the machine?

Jay Alammar. The illustrated transformer.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Wang Phil, and Samuel Weinbach. 2021. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch.

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, page 722–735, Berlin, Heidelberg. Springer-Verlag.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dzlab. [link].

Patrik Hrkút, Štefan Toth, Michal Ďuračík, Matej Meško, Emil Krsak, and Miroslava Mikušová. 2020. *Data Collection for Natural Language Processing Systems*, pages 60–70.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *CoRR*, abs/1808.07042.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *CoRR*, abs/2002.08910.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented SBERT: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 296–310, Online. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Lilian Weng. 2020. How to build an open-domain question answering system? *lilianweng.github.io*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022a. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022b. Linkbert: Pretraining language models with document links.

Vilém Zouhar, Marius Mosbach, Debanjali Biswas, and Dietrich Klakow. 2022. Artefact retrieval: Overview of nlp models with knowledge base access.