[CS591NR] Neural Networks: Neuroscience & Engineering Project Proposal Knowledge-Infused Language Model for Matching Chemical Ingredients and Cosmetic Products

Quoc Anh (Alan) Bui
Department of Computer Science
University of Massachusetts Amherst
qhbui@umass.edu

Anshumaan Chauhan
Department of Computer Science
University of Massachusetts Amherst
achauhan@umass.edu

I. INTRODUCTION

Cosmetics are an integral part of our daily lives and are used extensively for enhancing our appearance and maintaining personal hygiene. However, the safety and efficacy of cosmetic products are often a cause for concern due to the presence of various chemical ingredients. These ingredients can have adverse effects on the skin and overall health, making it essential to accurately match the ingredients to the intended cosmetic product.

In the recent years, studies have shown the ability of many powerful pretrained language models (PLMs) such as GPT, BERT, etc to learn the language representation which is helpful in downstream tasks such as *Masked Language Modeling* (MLM), *Named Entity Recognition* (NER), *Question Answering* (QA) and many more. The idea is to use unsupervised learning to train the model on a large text corpora which helps the model to learn the syntactical and semantic information. Furthermore the model is fine-tuned using a domain specific dataset for a downstream task. Supervised training is done in order to remove the bias and incorrect information learned in the initial training and to specialize the model for a more specific domain.

A major drawback of these models is the inability to keep up with the changing information. From a human perspective, if we do not learn about the new changes happening we cannot respond to any query or question related to it. To resolve this issue, external knowledge bases are used to provide the model with more contextual information about the input. In this work, we aim to explore the use of knowledge-infused language models for matching chemical ingredients to cosmetic products. The model will leverage the vast amounts of information available on the chemical properties and safety profiles of cosmetic ingredients to provide accurate and reliable ingredient-product matching.

In this research paper, we present a novel approach for ingredient matching in cosmetic products using a knowledge-infused language model. We will fine-tune a large language model that is pretrained on generic corpora in order to generate (or match) the label, i.e intended cosmetic use given a list of cosmetic ingredients. An additional step that we will do is to incorporate an external source of knowledge such as a *Knowledge Graph* (KG) into the model. The idea of knowledge graph is that it will additionally explore and introduce the relations between any two entities, i.e cosmetic ingredients to the model. We show how introducing the KG can affect the performance of the model on this downstream task, quantitatively and qualitatively.

II. LITERATURE SURVEY

In Natural Language Processing (NLP), the context of a sentence is used to resolve the ambiguity of a particular words or tokens, for example, bank can be related to either bank of a river or a financial bank. Support of additional knowledge which helps link the entities inside the given sentence using some relation has been an effective way to increase the performance of a Large Language Model (LLM).

Jiang et. al [1] compared the effect of various knowledge infusion techniques - addition of external knowledge as i) word embedding (ConceptNet) and ii) natural language sentences (ARC Corpus) to the baseline RoBERTa pretrained language model. Performance of the knowledge-infused models was seven percent better than the baseline model.

There are several domains, which do not have a structured dataset and therefore it is not feasible to fine-tune or train PLM in supervised manner. AgriBERT [3] was a agriculture specific PLM, a BERT model trained from scratch on a custom made dataset containing information related to food and agriculture. FoodOn Knowledge Graph formatted in OWL ontology was used to extract more information about the items, which were appended at the end of the question. They concluded that with increase in number of hops to find the relation between entities, Precision@1 increased but Mean Average Precision score decreased.

Rather than appending the external knowledge gained using KG, a Knowledge Injection component was used by Yan et. al [2]. This component takes the KG triplets (extracted from DBpedia) and positionally embed them into the input sentence using a Transformer and Embedding layer such that it abides by the natural language grammar, turning knowledge injection problem into a machine translation problem.

An ensemble of text based (takes the hypothesis and the premise as input) and graph based (input was the knowledge derived using ConceptNet in the form of graphs) models was introduced [4], using match-LSTM as the base model to solve the task of scientific textual entailment.

III. DATASET

A. Language Training Dataset

Due to unavailability of any structured (unlabeled) Language training textual dataset containing information about the products (especially product description), we will be making use of combination of Web Data Commons (Product domain dataset) and WikiText 103 (generic dataset).

B. Product Classification Dataset

We will be using Sephora Product dataset publicly available from Kaggle (23.26 MB). It consists of details such as Brand, Product Name, Ingredients, etc. for over 9000 products. Due to small size of the dataset it will be used for the purpose of fine-tuning the model parameters for the downstream task of mapping ingredients to the product label.

C. External Knowledge

There are some knowledge bases such as DBpedia, Yago, WordNet, ConceptNet and Freebase that contain a large amount of trusted and reliable information. Different knowledge base contain different types of information, for example, DBpedia is a generic knowledge base whereas ConceptNet consists of common sense knowledge. Product Classification does not require a lot of common sense reasoning to classify a product into a category based on its ingredients, therefore we will be using the generic knowledge base DBpedia.

IV. METHODOLOGY

A. Pretrained Large Language Models Fine-Tuning

Large Language Models are not familiar with the domain language semantics when trained on a general purpose corpora. Though they work well in many NLP tasks, they tend to hallucinate which results in an incorrect and biased predictions. Training LM on a dataset that consists of sentences/paragraphs related to the specific domain have worked well in biological and financial domains: BioBERT and FinBERT. For this purpose, we will train our pretrained BERT model on the product domain dataset, which will enhance the model's semantic and syntactical information. A mix of both WikiText and WDC will be used for domain specific training, which in total consists information about more than 26 million products. A standard approach for training the model on a large domain specific dataset is Masked Language Modeling - that is some fraction of the input text is removed from various positions and the model has to predict those masked words based on the context provided.

Once our model is trained on domain specific information, it is fine-tuned for the downstream task of product mapping/classification using Sephora dataset - consists of over 9000 entries (Figure 1). In the next section we will define how knowledge is injected along with the input text to ehance the model's performance.

B. External Knowledge Graph Injection

As presented in section 2, studies have shown how external knowledge injection have increased the model performance on several downstream NLP tasks. There are several ways of injecting knowledge gained from the external knowledge base such as using the triplets (head entity, tail entity, relation) as text appended at the end of the input, using a graph based model, conversion of the knowledge into a valid natural language grammar and positional embed it into the input and many more. Our proposed approach (Figure 2), takes in the input, which is passed to an entity recognizer. Entity recognizer has the task of recognizing different entities within the text and passing them on to external knowledge base. In the external knowledge base these entities are linked using different relationships, and the triplets are then provided as an output. These triplets will be converted into natural language texts, which will be appended at the end of original input.

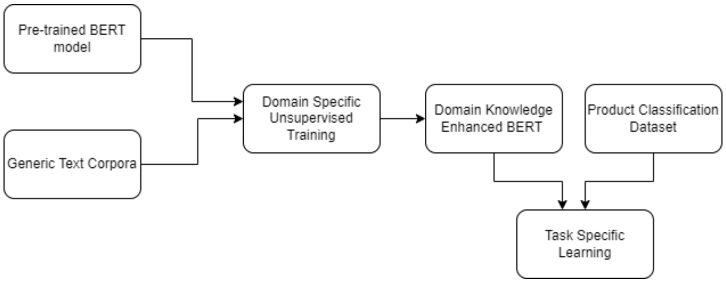
Pre-trained BERT model
Generic Text Corpora
Domain Specific Unsupervised Training
Domain Knowledge Enhanced BERT
Product Classification Dataset
Task Specific Learning

Figure 1. LLM task-specific fine-tuning

V. EVALUATION

In Machine Learning, when we are dealing with classification problems, *categorical cross-entropy* loss is used as a loss function during model training. The performance metric used during testing the model with other baseline architectures will be *Accuracy*.

$$Accuracy = \frac{\sum \text{Correctly classified labels}}{\sum \text{All classified labels}} \quad (1)$$

Accuracy is a measure of how many correct predictions/ mappings are made.

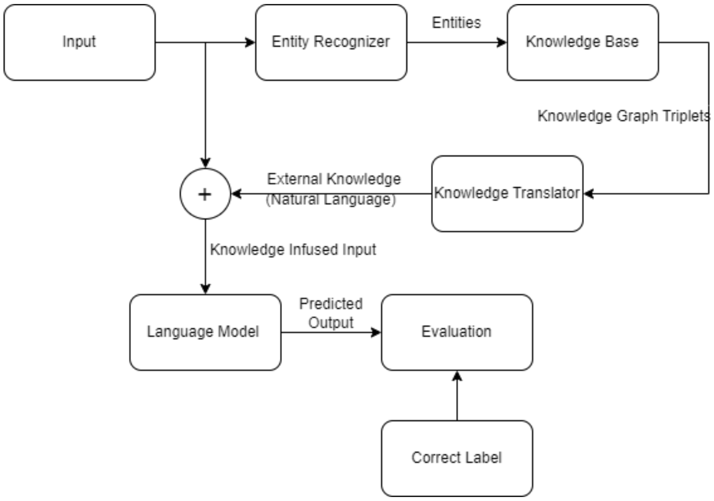

Figure 2. External Knowledge Injection

In the first trial, the model will perform on the test dataset without the help of the KG injection. The advice/ context will be introduced to the model via KG in the second trial. We wish to establish the difference and conclude the performances before and after injecting external knowledge.

References

[1] Jiang, Yichuan, and Hyan Huang. Analysis and improvement of external knowledge usage in machine multi-choice reading comprehension tasks. *2020 2nd International Conference on Machine International Conference on Information Knowledge Management*, 2021. 1

[2] Yan Ruiqing and et la. A general method for transferring explicit knowledge into language model pretraining. *Security and Communication Networks*, 2021. 1

[3] Rezayi Saed and et al. Agribert: Knowledge-infused agricultural langauge models for matching food and nutrition. *IJCAI*, 2022. 1

[4] Wang Xiaoyan and et la. Improving natural language inference using external knowledge in the science questions domain. *Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01*, 2019. 2