Inductive Transfer Learning for Detection of Well-formed Natural Language Search Queries

by

Bakhtiyar Syed, Vijaysaradhi Indurthi, Manish Gupta, Manish Shrivastava, Vasudeva Varma

in

41st European Conference on Information Retrieval
(*ECIR-2019*)

Cologne, Germany

Report No: IIIT/TR/2019/-1

Centre for Search and Information Extraction Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2019

Inductive Transfer Learning for Detection of Well-formed Natural Language Search Queries

Bakhtiyar Syed*[1], Vijayasaradhi Indurthi*[1], Manish Gupta[1,2], Manish Shrivastava[1], and Vasudeva Varma[1]

[1] IIIT Hyderabad
{syed.b, vijaya.saradhi}@research.iiit.ac.in, {manish.gupta, m.shrivastava, vv}@iiit.ac.in
[2] Microsoft
gmanish@microsoft.com

Abstract. Users have been trained to type keyword queries on search engines. However, recently there has been a significant rise in the number of verbose queries. Often times such queries are not well-formed. The lack of well-formedness in the query might adversely impact the downstream pipeline which processes these queries. A well-formed natural language question as a search query aids heavily in reducing errors in downstream tasks and further helps in improved query understanding. In this paper, we employ an inductive transfer learning technique by fine-tuning a pretrained language model to identify whether a search query is a well-formed natural language question or not. We show that our model trained on a recently released benchmark dataset spanning 25,100 queries gives an accuracy of 75.03% thereby improving by ∼5 absolute percentage points over the state-of-the-art.

1 Introduction

Traditionally users have been trained to put up keyword queries on search engines mainly because search has been traditionally driven by "unigram match". However, recently, with the increasing popularity of voice based search, verbose queries have become quite popular [9]. Also, in-vogue deep learning algorithms have enabled search engines to process such verbose natural language (NL) queries effectively. But not all verbose queries from users exhibit proper structure. Such queries often lack a coherent structure and may sometimes violate grammar rules, thus mandating tailor-made processing [2, 4, 11, 14]. This makes it challenging for NL Processing (NLP) tools trained on formal text to extract the relevant information required to understand the user's intention behind the query [1].

Identifying whether a search query is well-formed [8] is an important task which also aids in various downstream tasks like understanding the user's intent (in case of personal assistants and chatbots [15, 18]) and generating better related query suggestions in search engines.

* The authors contributed equally.

A possible approach for improving the accuracy of downstream processing while still using malformed queries is to train models using labeled data with malformed queries. But this approach has two main drawbacks. First, getting annotations to generate such training data which captures all possible malformed variants can be quite expensive. Second, since there is frequent change in the nature and domain of these queries [3, 12, 17], any model which is trained on these queries will drift fairly quickly. Another possible approach is to use grammars for identification of query well-formedness. Ideally, grammars such as the grammar on English resource [5] should be able to identify whether a query is a well-formed question or not easily. But in practice, such grammars are very precise and are not able to accurately parse more than half of web search queries.

The idea of using inductive transfer learning in natural language processing, akin to allied areas in computer vision like object detection, image segmentation, etc. has been garnering attention. Most of the current research focused on training deep learning models from scratch require huge volumes of training data and are also computationally expensive. Recent advancements in training and fine-tuning language models (LMs) are being used for a variety of NLP applications and have shown significant promise, primarily in text classification tasks [10]. In this work, we show that inductive transfer learning is greatly beneficial in identifying well-formed natural language questions. We also perform ablation studies to show the effectiveness of each of the modules used in the inductive transfer learning technique. Our experiments show an accuracy of 75.03% on the benchmark dataset for the task improving by $\sim$5 absolute percentage points over the state-of-the-art method [8].

2 Related Work

Faruqui and Das [8] introduce the task and provide a coherent understanding of why many of the current techniques are not suitable for detecting well-formed questions. They combine various NLP features like word, character and Part-of-Speech (POS) n-grams with a simple neural network for the task. We show improvements over their method in Section 5. Attempts to identify well-formed questions by parsing them using grammar like the English resource grammar [5] are not very effective as the grammar is highly precise and fails to parse significant fraction of the web queries. Two other related tasks are Grammatical Error Prediction (GEP) and Correction (GEC) [18, 19]. While GEP is simply the task of classifying whether a given sentence is grammatical, GEC is a more complex task which involves identifying parts of ungrammatical text and correcting the same to produce grammatically correct text. Although our work is similar to GEP, past research has explored GEP for fully formed sentences which may not have a search intent. Thus, unlike GEP/GEC, we focus on well-formedness check for web search queries which contain a specific user intent.

3 Proposed Approach: Inductive Transfer Learning (ITL)

In this section, we first define the problem formally and then discuss three important phases of our inductive transfer learning approach in detail.

The architecture diagram of the proposed approach is illustrated in Figure 1.

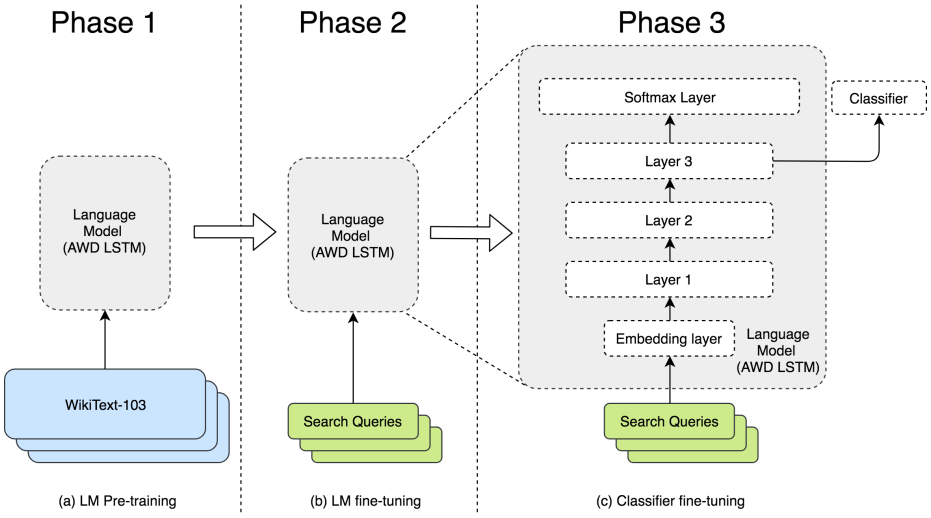

Fig. 1. Inductive Transfer Learning mechanism to identify well-formed search queries

Query Well-formedness Detection Problem: Given a query q, we intend to learn the label C which describes whether the query is a well-formed natural language question or not. We model this task as a binary classification task: C=1 indicates that the query is well-formed while C =0 indicates non-well-formedness.

The ULMFiT Architecture: Previous attempts to use inductive transfer through language modeling have resulted in limited success for NLP tasks [6, 16]. However, Howard and Ruder [10] showed that if language models (LMs) are fine-tuned correctly, they would not overfit to small datasets and would enable robust inductive transfer learning. The neural architecture is called the Universal Language Model Fine Tuning architecture. They also proposed novel techniques which prevent catastrophic forgetting during training of the language model. We adapt the ULMFiT model for our inductive transfer learning approach and show that inductive transfer learning is greatly beneficial for identifying well-formed natural language search queries.

The AWD-LSTM model: Our inductive transfer learning mechanism utilizes the state-of-the-art Averaged-SGD Weight-Dropped Long Short Term Memory (AWD-LSTM) networks [13]. It is a variant of the simple LSTM with no shortcut connections, no attention or any other advanced mechanisms, with the same hyperparameters as in typical LSTMs, and no additions other than tuned drop-

connect hyperparameters. We use AWD-LSTMs since they have been shown to be effective in learning lower-perplexity language models.

Three Stages of the proposed ITL Framework: The proposed ITL framework for query well-formedness check involves these three important phases:

1. **General Domain Pretraining**: The first phase involves pretraining a language model on a huge English corpus. In our case, we use the pretrained language model trained on the released Wikitext-103 [13] dataset which consists of 103 Million unique words and 28,595 preprocessed Wikipedia articles. This helps the model to learn the general language dependencies and is the first step before fine-tuning which targets task-specific data.
2. **Language Model Fine-tuning for the Target Task**: The data used for the target task is usually from a specific distribution (as compared to the general distribution in the large corpus used in the previous phase). Clearly, it is essential information for the language model – no matter how diverse the general domain data in the earlier pretraining step is. Hence, in this phase, we use task-specific data to fine-tuning our language model in an unsupervised manner. As proposed in [10], our fine-tuning involves discriminative fine-tuning and slanted triangular learning rates to combat the catastrophic forgetting language models exhibited in previous works [16, 6] which used language models for fine-tuning.
   Discriminative Fine-tuning (DFT): Instead of keeping the same learning rate for all the layers of the AWD-LSTM, a different learning rate is used for tuning the three different layers. The intuition behind this is that since each of the layers represent a different kind of information [20], they must be fine-tuned to different extents.
   Slanted Triangular Learning Rates (STLR): Using the same learning rate is not the best way to enable the model to converge to a suitable region of the parameter space. Thus we adapt the slanted triangular learning rate [10] which first increases the learning rate and then linearly decays it as the number of training samples increases.
3. **Classifier Fine-tuning for the Target Task**: The weights that we obtain from the second phase are fine-tuned by keeping the same upstream architecture, but also appending 2 fully connected layers for the final classification with the last layer predicting the well-formedness rating. In this phase, we adapt the *gradual unfreezing heuristic* [10] for our task.
   Gradual Unfreezing (GU): All layers are not fine-tuned at the same time, instead the model is gradually unfrozen starting from the last layer, as it contains the least general knowledge [20]. The last layer is first unfrozen and fine-tuned for one epoch. Subsequently, the next frozen layer is unfrozen and all unfrozen layers are fine-tuned. This is repeated until all layers are fine-tuned until convergence is reached.

4 Dataset

For our experiments, we use the recently released benchmark dataset for well-formed natural language questions [8]. It contains a total of 25,100 questions,

each labeled with a rating (between 0 and 1) of the query being well-formed. The authors collected these questions by utilizing questions asked by users on WikiAnswers[3], originally published as the Parallex corpus [7]. The compiled dataset primarily contains well-formed questions as queries along with typical constructs of search queries.

A query is annotated as well-formed if the supplied query is *grammatical* in nature, has *perfect spellings* and is an *explicit question*. For each search query, the average of the five scores (over each of the annotator's ratings) is calculated and then documented as the final rating R which indicates the degree of its well-formedness. As suggested in [8], the query is considered well-formed if $R \geq 0.8$ for the query[4]. In our experiments, we make use of the standard train-dev-test split supplied by Faruqui and Das [8], which consists of 17500 training, 3750 development and 3850 test queries. A few queries with their corresponding labels from the dataset are shown in Table 1.

Table 1. Examples from the benchmark well-formedness dataset from [8]

Example	**Well-formedness Rating**
Which form of government is still in place in greece ?	1.0
One of Mussolini 's goals ?	0.0
How many leagues in a mile in the mexican term ?	0.4
What is the scotlands longest river ?	0.2
How do you get rid of browsing history ?	0.8

5 Experiments

Baselines: We compare our proposed method with the following baselines.

- Majority Class Prediction: Classify all queries into the majority class in the test set.
- Question Word Classifier: If the query starts with an interrogative word, classify it as being well-formed.
- Word Bi-Directional LSTM (BiLSTM) Classifier: Use a Bi-LSTM for classification which takes as input a one-hot vector of the input words of the query, and classifies using a softmax for binary classification.
- Faruqui and Das [8] propose a 2 hidden layer neural architecture with rectified linear unit (ReLU) activations and a final softmax layer for the predictions. For the input, the authors extract word, character and Part-of-Speech (POS) *n-grams*: word-1,2; char-3,4 grams as the lexical features and POS-1,2,3 grams for the syntactic features to form the n-gram embeddings via concatenation.

[3] http://www.answers.com/Q/

[4] A rating greater than or equal to 0.8 ensures at least 4 out of 5 annotators marked the query as well-formed.

Table 2. Comparison of Various Classifiers and Ablation Study for the ITL Model

Model	Accuracy (%)
Question Word Classifier	54.9
Majority Class Prediction	61.5
Word BiLSTM Classifier	65.8
word-1,2 char-3,4 grams [8]	66.9
word-1,2 POS-1,2,3 grams [8]	70.7
word-1,2 char-3,4 POS-1,2,3 grams [8]	70.2
.....	.....
(Inductive Transfer Learning)	
No pretraining with WikiText-103	68.2
No LM fine-tuning	72.8
Fine-tuning without DFT and STLR	73.0
No gradual unfreezing	72.4
All (Pre-train + Fine-tune with DFT and STLR + Gradual unfreezing)	**75.0**

Hyper-parameter Settings: As suggested in [10], we use the AWD-LSTM language model with 3 layers, 1150 hidden activations per layer and an embedding size of 400. The hidden layer of the classifier is of size 50. A batch size of 30 is used to train the model. The LM and classifier fine-tuning is done with a base learning rate of 0.004 and 0.01 respectively. All experiments are on the standard train-dev-test split as proposed in [8] with the classification results reported on the 3850 sized test data.

Results and Analysis: Table 2 shows the performance of our ITL model as compared with various baselines and the state-of-the-art method [8]. The overall ITL model has an accuracy of 75.03% improving significantly over the previous state-of-the-art (feature-engineered solution of [8]). To assess the impact of each of the three steps involved in ITL, we perform an ablation study as follows.

- **No pretraining**: Train the model without the pretraining step.
- **No LM fine-tuning**: Phase 2 is ignored.
- **Fine-tuning without DFT and STLR**: LM fine-tuning without discriminative fine-tuning and without Slanted Triangular Learning Rates.
- **No gradual unfreezing**: No gradual unfreezing during classifier fine-tuning.

As expected, from Table 2, we observe that not fine-tuning the LM on the target task results in a worse performance versus fine-tuning. Using DFT and STLR is beneficial. Gradual unfreezing helps in increasing the performance. All the three steps in the fine-tuning process contribute towards improving accuracy.

6 Conclusions

In this work, we showed that the idea of using inductive transfer learning by fine-tuning language models aids in identifying whether search queries are well-formed natural language questions. On a large dataset of 25,100 questions, we showed that our method beats the baselines with a significant margin. In the future, we plan to explore the "accuracy versus labeled dataset size" tradeoff for this approach across multiple resource-poor languages.

References

1. Baeza-Yates, R., Calderón-Benavides, L., González-Caro, C.: The intention behind web queries. In: International Symposium on String Processing and Information Retrieval. pp. 98–109. Springer (2006)
2. Barr, C., Jones, R., Regelson, M.: The linguistic structure of english web-search queries. In: Proceedings of the conference on empirical methods in natural language processing. pp. 1021–1030. Association for Computational Linguistics (2008)
3. Bawa, M., Bayardo Jr, R.J., Rajagopalan, S., Shekita, E.J.: Make it fresh, make it quick: searching a network of personal webservers. In: Proceedings of the 12th international conference on World Wide Web. pp. 577–586. ACM (2003)
4. Bergsma, S., Wang, Q.I.: Learning noun phrase query segmentation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (2007)
5. Copestake, A.A., Flickinger, D.: An open source grammar development environment and broad-coverage english grammar using hpsg. In: LREC. pp. 591–600. Athens, Greece (2000)
6. Dai, A.M., Le, Q.V.: Semi-supervised sequence learning. In: Advances in neural information processing systems. pp. 3079–3087 (2015)
7. Fader, A., Zettlemoyer, L., Etzioni, O.: Paraphrase-driven learning for open question answering. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1608–1618 (2013)
8. Faruqui, M., Das, D.: Identifying well-formed natural language questions. In: EMNLP. p. To Appear (2018)
9. Gupta, M., Bendersky, M., et al.: Information retrieval with verbose queries. Foundations and Trends® in Information Retrieval **9**(3-4), 209–354 (2015)
10. Howard, J., Ruder, S.: 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). vol. 1, pp. 328–339 (2018)
11. Manshadi, M., Li, X.: Semantic tagging of web search queries. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 861–869. Association for Computational Linguistics (2009)
12. Markatos, E.P.: On caching search engine query results. Computer Communications **24**(2), 137–143 (2001)
13. Merity, S., Keskar, N.S., Socher, R.: Regularizing and optimizing lstm language models. arXiv preprint arXiv:1708.02182 (2017)
14. Mishra, N., Saha Roy, R., Ganguly, N., Laxman, S., Choudhury, M.: Unsupervised query segmentation using only query logs. In: Proceedings of the 20th international conference companion on World wide web. pp. 91–92. ACM (2011)
15. Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., Vanderwende, L.: Generating natural questions about an image. arXiv preprint arXiv:1603.06059 (2016)
16. Mou, L., Meng, Z., Yan, R., Li, G., Xu, Y., Zhang, L., Jin, Z.: How transferable are neural networks in nlp applications? arXiv preprint arXiv:1603.06111 (2016)
17. Roy, R.S., Choudhury, M., Bali, K.: Are web search queries an evolving protolanguage? In: The Evolution Of Language, pp. 304–311. World Scientific (2012)
18. Yang, J., Hauff, C., Bozzon, A., Houben, G.J.: Asking the right question in collaborative q&a systems. In: Proceedings of the 25th ACM conference on Hypertext and social media. pp. 179–189. ACM (2014)

19. Yannakoudakis, H., Rei, M., Andersen, Ø.E., Yuan, Z.: Neural sequence-labelling models for grammatical error correction. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP. pp. 2795–2806 (2017)
20. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems. pp. 3320–3328 (2014)