Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, June : 2023

Neural Duplicate Question Detection without Labelled Training Data

[1] **A. RAKESH.**[2] **B.SUMANTH,** [3] **CH.AJAY**, B-TECH student, Department of **Electronics Communication Engineering,** Vignana Bharathi institute of technology, Hyderabad,
B. Sri Krishna, Assistant professor, Department of **Electronics Communication Engineering,** Vignana Bharathi institute of technology, Hyderabad,
Alpularakesh002@gmail.com balnesumanth@gmail.com ajayrudhra1383@gmail.com

Abstract:

Large quantities of expensively acquired labelled question pairs are needed for the supervised training of neural networks for duplicate question identification in community question answering (cQA). Recent efforts have frequently used alternate techniques to reduce this cost, such as adversarial domain adaptation. In this paper, we suggest two unique approaches: (1) automatic question duplication; and (2) weak supervision based on the body and title of the question. We demonstrate that both can perform better even when they don't need labelled data. We present thorough comparisons of well-liked training approaches, which offer crucial insights on how to "best" train models in various contexts. We demonstrate that the increased ability to use greater quantities of unlabeled data from cQA forums makes our suggested approaches more efficient in many situations.

Keywords: neural networks, cQA, Neural Duplicate Question.

I. INTRODUCTION:

It is critical to automatically detect duplicate questions in community question-answering (cQA) forums so that users can avoid asking the same query more than once and more quickly locate pre-existing inquiries and answers. For the supervised training of neural approaches for duplication detection, large numbers of labelled question pairs—that is, labelled pairs of duplicate questions that can be answered using the same data—are frequently required.

In reality, obtaining such data is frequently challenging due to the prodigious amount of manual labour necessary for annotation. As a result, many cQA forums lack enough labelled data for supervised neural model training.

As a result, newer studies have employed various training techniques. This includes adversarial domain transfer, semi-supervised training, and poor supervision with question-answer pairs. These algorithms, however, rely on huge amounts of labelled data, which may include a large number of question-answer pairings or thousands of duplicate questions.

Unsupervised techniques also rely on encoder-decoder architectures, which constrain model topologies and typically yield subpar results in comparison to supervised training; instead, they must be paired with sophisticated features to yield cutting-edge results. For the several cQA forums' effective duplicate question detection models, without labelled duplicates As a result, we need additional techniques that are as effective as supervised in-domain training without requiring any annotations.

In this work, we suggest two unique approaches: (1) automatic duplicate question generation (DQG) (WS-TB); and (2) poor oversight of the title-body combinations in situations where we can only access unlabeled questions. Due to the fact that question bodies frequently contain additional significant information that is not included in the title, we hypothesise that bodies and titles have similar qualities as duplicate questions. They effectively describe the same question, for instance, but are just marginally repetitive. As a result, we may train our models using the data from titles and bodies as well as their relationships.

In this research, we compare a wide range of training methods and evaluate well-known duplicate detection and question retrieval models such as RCNN and BiLSTM: supervised training, adversarial domain transfer, question-answer pairs for poor supervision, DQG, WSTB, and unsupervised training.

We exhibit that:

1. Information from the title-body is especially helpful for training models. When there are more DQG, WS-TB and unlabeled questions, outperform adversarial domain transfer from comparable source domains by an average of more than 5.8pp. Because there are frequently fewer labelled question duplicates, WS-TB and DQG can occasionally perform better than supervised training.
2. As DQG is domain-translatable, question generation models can be utilised to create duplicates that are appropriate for training models in new target domains.
3. Our training techniques work well for optimising more modern models like BERT (Devlin et al., 2018a)
4. Without direct answer supervision, WS-TB can likewise be utilised to train cQA answer selection models. This demonstrates that our methods can have an influence on jobs that are linked to duplicate question detection and beyond.

II. LITERATURE SURVEY

The capacity of deep neural networks (DNNs) to automatically engineer features is a significant benefit when employing them for text applications. Sadly, DNNs frequently require a large amount of training data, especially for high-level semantic applications like community question answering (cQA). Using multitask learning to learn the target DNN along with two auxiliary tasks, we address the issue of data scarcity in this study. We take advantage of the strong semantic link between the comments we choose that are pertinent to (i) I-Know inquiries and (ii) Forum queries. This makes it possible for comments and new and old queries to be represented globally. A SemEval challenge dataset used in our model's experiments for cQA shows a 20% relative improvement over conventional DNNs. [1]

The capacity of deep neural networks (DNNs) to automatically engineer features is a fundamental benefit when employing them for text applications. Sadly, DNNs frequently require a large amount of training data, especially for complex semantic applications like responding to community questions (cQA). By using multitask learning to learn the target DNN along with two auxiliary tasks, we are able to address the issue of data scarcity in this work. The remarks we choose that are pertinent to new inquiries and forum queries, respectively, take advantage of the close semantic relationship between them. This enables comments and both new and old inquiries to be universally reflected. Our model's cQA tests on a SemEval challenge dataset demonstrate a 20% relative improvement over traditional DNNs. [2]

Sentence Abstract summarization attempts to retain the sense of a given sentence while producing a shorter version of it. We propose a conditional RNN, a recurrent neural network (RNN), that produces a summary of an input text. The conditioning ensures that the decoder concentrates on the pertinent input words at each stage of development using a special convolutional attention-based encoder. Our model is simple to train end-to-end on big data sets and only uses learned features. Our tests demonstrate that the model performs competitively on the DUC-2004 shared task while dramatically outperforming the previously proposed state-of-the-art technique on the Gig word corpus. [4]

The new paradigm we propose for encoding languages is BERT, or Bidirectional Encoder Representations from Transformers. In contrast to modern language representation models, BERT conditions both left and right context in all layers at the same time to train deep bidirectional representations from unlabeled text. So, the pre-trained BERT model can be improved by adding one more output layer without making big changes to the system's architecture. This makes it possible to make cutting-edge models for tasks like language inference and question answering. Both experimentally and conceptually, BERT makes sense. [6]

We look at the challenge of creating question-answer pairs from Wikipedia articles that cover more than one sentence's worth of information. We suggest a neural network strategy that uses a cutting-edge gating technique to include coreference knowledge. We find that the linguistic knowledge supplied by the coreference representation greatly aids question generation when compared to models that merely use sentence-level information. This produces models that outperform the state-of-the-art at this time. Using the top 10,000 Wikipedia articles, we apply our system, which consists of an answer span extraction technique and a passage-level QG system, to create a corpus of more than a million question-answer pairings. We also conducted a qualitative examination of this substantial corpus of texts created by Wikipedia. [7]

III. PROPOSED METHOD

The drawback of the current approaches is that they call for the transfer of labelled question duplicates, acceptable responses, target domains, and similar sources. Another option is to use unsupervised training inside of an encoder-decoder framework. However, this puts a lot of restrictions on the network architecture, like the need to encode each question separately.

These shortcomings, such as the need for labelled data and architectural restrictions, are not present in our suggested methods.

Duplicate question generation (DQG):

Creates new question titles from the contents of the questions, which we then treat as copies of the original titles. Figure 2 shows our overall strategy. We could use this method to train duplicate detection models for all cQA forums with a large enough number of unlabelled title-body pairings to get a good QG model, since DQG doesn't need data to be labelled.

Weak supervision with title-body pairs (WSTB):

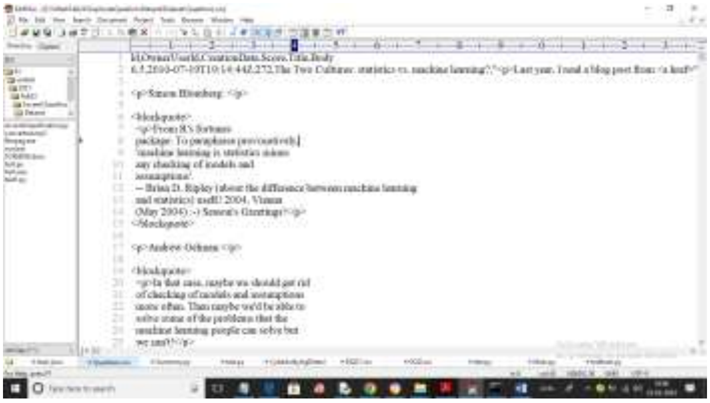
Takes the DQG premise one step farther. We might also train duplicate detection models on this data without first creating questions if question body and question titles have comparable features as duplicates. Due to the fact that it doesn't need a separate question generating model, this strategy makes gathering training data significantly simpler.

IV. RESULT

In this paper introducing two novel techniques which can identify duplicate questions without any labelling. In first technique author applying TRANSFORMER algorithm to generate NEW QUESTION from TITLE and then this question will check with original TITLES as we know user will write same questions with different words and meaning so transformer will generate such duplicate questions. Both transformer question and original question will be get similarity ranked and if ranked is higher then that question will be marked as DUPLICATE and this technique is called as DUPLICATE QUESTION GENERATOR. Generated labels will be trained with RCNN and BILSTM to calculate correct prediction AUC value and if question predicted correctly as duplicate or non-duplicate then AUC value will be high

Second technique will take both TITLE and BODY to generate new question and this newly generated question will check with original title to get ranking and if ranking is high then duplicated question will be detected. Both body and title may give high similarity of duplicate question so its AUC value will be high compare to DQG first technique. This technique is called as WSTB (weak supervise Title body pair)

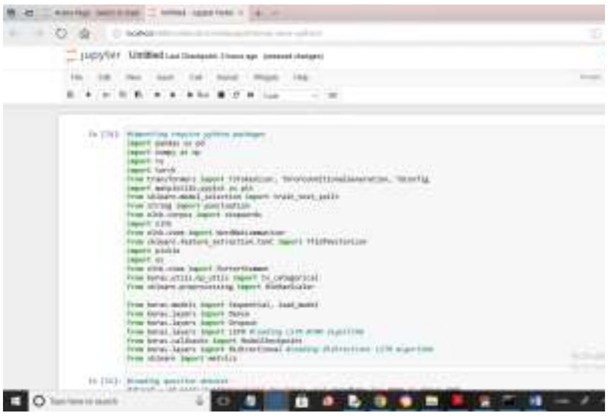
To train both algorithms we have downloaded QUESTIONS dataset from KAGGLE which contains question title and body and by using this dataset we have evaluated performance of both algorithms and below screen showing dataset details. From same dataset author has separated questions as “Ask Ubuntu, Android and many more” but it’s difficult to separate so we used entire dataset



```
Id,OwnerUserId,CreationDate,Score,Title,Body
6,5,2010-07-19T19:14:44Z,272,The Two Cultures: statistics vs. machine learning?,"<p>Last year, I read a blog post from <a href=""

<p>Simon Blomberg: </p>

<blockquote>
  <p>From R's fortunes
  package: To paraphrase provocatively,
  'machine learning is statistics minus
  any checking of models and
  assumptions'.
  -- Brian D. Ripley (about the difference between machine learning
  and statistics) useR! 2004, Vienna
  (May 2004) :-) Season's Greetings!</p>
</blockquote>

<p>Andrew Gelman: </p>

<blockquote>
  <p>In that case, maybe we should get rid
  of checking of models and assumptions
  more often. Then maybe we'd be able to
  solve some of the problems that the
  machine learning people can solve but
  we can't!</p>
```

In above dataset screen first row contains dataset column names and remaining rows contains dataset values.

We have coded this project using JUPYTER NOTEBOOK and below are the code and output screen with blue colour comments

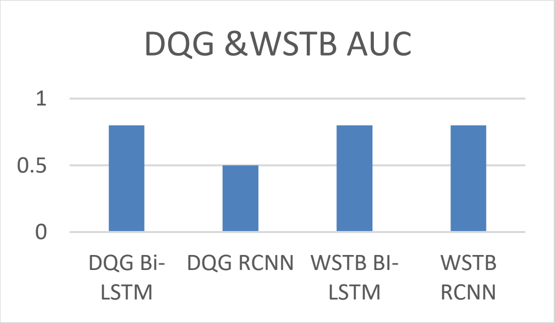


```
In [76]: #importing require python packages
         import pandas as pd
         import numpy as np
         import re
         import torch
         from transformers import T5Tokenizer, T5ForConditionalGeneration, T5Config
         import matplotlib.pyplot as plt
         from sklearn.model_selection import train_test_split
         from string import punctuation
         from nltk.corpus import stopwords
         import nltk
         from nltk.stem import WordNetLemmatizer
         from sklearn.feature_extraction.text import TfidfVectorizer
         import pickle
         import os
         from nltk.stem import PorterStemmer
         from keras.utils.np_utils import to_categorical
         from sklearn.preprocessing import MinMaxScaler

         from keras.models import Sequential, load_model
         from keras.layers import Dense
         from keras.layers import Dropout
         from keras.layers import LSTM #loading LSTM RCNN algorithm
         from keras.callbacks import ModelCheckpoint
         from keras.layers import Bidirectional #loading Bidirectional LSTM algorithm
         from sklearn import metrics

In [16]: #loading question dataset
```

In above screen we are loading all require python classes and packages and you can read blue colour comments to know about coding

In above screen DQG dataset we are training with RCNN LSTM algorithm and we got its AUC values as 0.5% which you can see in blue colour text

In above screen we are training DQG data with BI-LSTM algorithm and we got its AUC value in blue colour text as 0.85%



In above graph x-axis represents algorithm names and y-axis represents AUC values and in both techniques WSTB got high AUC (area under curve)

	Algorithm Name	RCNN AUC	Bi-LSTM AUC
0	DQG	0.500000	0.851852
1	Purpose WSTB	0.950549	0.963370

In above screen in tabular format we can see AUC values for both algorithms

V. CONCLUSION

Without labelled training data, we trained duplicate question detection algorithms for our investigation. This can be useful for many cQA forums that don't have enough duplicate questions with annotations or question-answer pairings to employ the most recent training methods. Our two unique ways to duplicate weak supervision and question generation using title-body pairs can use more data during training because they only use the title-body information of unlabelled questions. Even though both supervised training and using more unlabelled questions are very effective when using the same number of training samples as other methods, our research shows that we can do even better by using more unlabelled questions. Additionally, we have shown that answer selection models may be trained without direct answer supervision using weak supervision with title-body pairings. This demonstrates that our research could potentially help with a far larger range of related activities than just question duplication.

REFERENCES

1. Daniele Bonadiman, Antonio Uva, and Alessandro Moschitti. 2017. Effective shared representations with multitask learning for community question answering. In Proceedings of the 15th Conference of the European

Chapter of the Association for Computational Linguistics (EACL 2017), pages 726–732.

2. Xin Cao, Gao Cong, Bin Cui, Christian S. Jensen, and Quan Yuan. 2012. Approaches to exploring category information for question retrieval in community question-answer archives. ACM Transactions on Information Systems (TOIS), 30(2):7:1–7:38.
3. Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016), pages 93–98.
4. Giovanni Da San Martino, Alberto Barron Cede ´ no, Sal- ˜ vatore Romeo, Antonio Uva, and Alessandro Moschitti. 2016. Learning to re-rank questions in community question answering using advanced features. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM 2016), pages 1997–2000.
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint, abs/1810.04805.
6. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
7. Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), pages 1907–1917.