IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

A Structural, Time Aware, Coordinated Tag Generation Based on Transformer Network

Prof. Shwetha G R[1], Snehith Prasad C H[2], Shiva Prasad C[3]
Assistant Professor, Department of Information Science and Engineering[1]
Students, Department of Information Science and Engineering[2,3]
S J C Institute of Technology, Chickballapur, India

Abstract: *The content quality of shared knowledge in Stack Overflow (SO) is critical in supporting software developers with their programming problems. Thus, it allows its users to suggest editing the software to improve the quality of a post. However, existing all research shows that many suggested edits in SO are rejected due to undesired contents or violating editing guidelines. Such a scenario frustrates or demotivates users who would like to conduct good-quality edits. we propose Semantically Tag and Score Recommendation, with the use of the deep learning-based approach that automatically recommends tags or grades or scores through learning the semantics of both tags, score, grade and questions in such software CQA. First, word embedding is employed to convert text information to high-dimension vectors for better representing questions and tags. Second, a Multitasking, the core modules of Semantically Tag and Score Recommendation, is designed to capture short and long semantics. Third, the learned semantic vectors are fed into a gradient descent-based algorithm for classification.*

Keywords: Tag Generation

I. INTRODUCTION

Structural time-aware coordinated tag generation based on transformer network is a technique used in natural language processing (NLP) to generate coordinated tags for a given sentence, while taking into account the underlying temporal structure of the sentence. This technique involves the use of a transformer network, which is a deep learning architecture that has been widely used in NLP for various tasks such as language translation, text summarization, and sentiment analysis. The coordinated tags generated by this technique are designed to provide a more complete and accurate representation of the sentence, by capturing both the syntactic and temporal relationships between the different elements of the sentence. This is achieved by incorporating information about the sentence's structure and context, as well as the timing of events and actions described in the sentence.

The transformer network used in this technique consists of multiple layers of self-attention mechanisms, which enable the model to focus on different parts of the input sentence at different times. The network is trained on a large corpus of text, using supervised learning techniques, to learn how to generate coordinated tags for a given sentence based on its underlying structure and temporal relationships. Overall, the structural time-aware coordinated tag generation based on transformer network is an advanced technique that offers significant benefits for natural language processing tasks, particularly those involving complex sentences with multiple clauses and temporal dependencies. It has the potential to greatly enhance the accuracy and effectiveness of various NLP applications, including text classification, information retrieval, and question answering systems.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

II. ARCHITECTURE

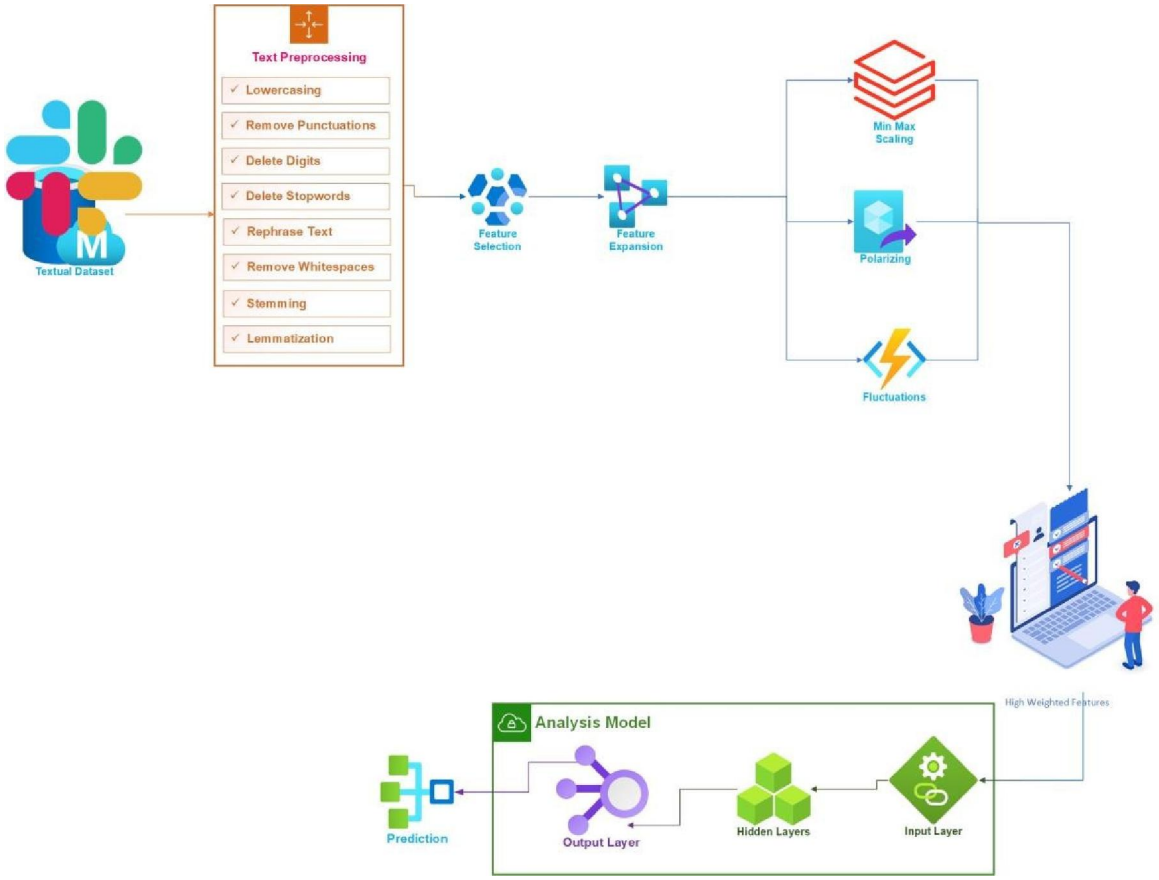

Figure: Proposed System Architecture

III. PROPOSED ALGORITHM

Code Bidirectional Encoder Representation Transformers (CodeBERT)
Algorithm Bi-directional Long Short-Term
Memory (LSTM) Algorithm (BiLSTM)

The Bidirectional Encoder Representations from Transformers (BERT) model is a pre- trained language model that has gained a lot of popularity in recent years due to its impressive performance on a wide range of natural language processing tasks. Here are some of the advantages of using BERT: Over-all. Improved Contextual Representation: BERT is a bi-directional model that uses a transformer architecture to encode the context of a given word. This means that the model considers the context of the entire sentence when generating word embeddings, which leads to a much more accurate representation of the sentence and its meaning. Pre-Trained Model: BERT is pre-trained on a large corpus of text, which means that it has already learned a lot about the structure and context of natural language. BiLSTM (Bidirectional Long Short-Term Memory) is a type of recurrent neural network that is widely used in natural language processing (NLP) and speech recognition tasks. Here are some of the advantages of using BiLSTM Better Contextual Understanding: BiLSTM models are able to capture both the past and future context of a given word or sequence, allowing for a more comprehensive understanding of the text. noisy data, such as punctuation, special characters, and HTML tags. This can be done using regular expressions or other text processing techniques. Sentence Tokenization: The dataset should be

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

Advantages of Proposed Algorithm

This is particularly useful in tasks such as sentiment analysis, where the meaning of a sentence can depend on the words that come before and after it. Reduced Vanishing Gradient Problem: LSTM (Long Short-Term Memory) networks, on which BiLSTM is based, were designed to overcome the vanishing gradient problem that occurs in traditional recurrent neural networks. This makes BiLSTM more effective at handling long sequences of text, where traditional models may struggle to capture long-term dependencies.

IV. PROJECT MODULES

Module 1: Data Preparation

Data preparation is a crucial step in any machine learning task, and it is particularly important for structural time-aware coordinated tag generation based on transformer network. Here are some of the key steps involved in data preparation for this task: Data Collection: The first step is to collect a dataset of sentences that will be used to train and test the model. The dataset should be representative of the type of sentences that the model will be expected to handle. Data Cleaning: The dataset should be cleaned to remove any irrelevant or tokenized into individual sentences using a natural language processing library, such as NLTK or SpaCy. Word Tokenization: Each sentence should be further tokenized into individual words using the same natural language processing library. POS Tagging: Each word in the sentence should be tagged with its part of speech (POS) using a POS tagger, such as the Stanford POS tagger. Dependency Parsing: The dataset should be parsed to identify the dependencies between the words in each sentence using a dependency parser, such as the Stanford dependency parser. Coordinated Tagging: Each sentence should be annotated with coordinated tags, which indicate the relationships between different parts of the sentence. This can be done manually by human annotators, or using a tool such as the Stanford CoreNLP. Data Splitting: Finally, the dataset should be split into training, validation, and test sets, with a sufficient number of examples in each set to ensure that the model can generalize well to new data. Overall, data preparation is a critical step in ensuring the accuracy and effectiveness of a structural time-aware coordinated tag generation based on transformer network model. Properly prepared data can help the model to learn the underlying structure of natural language sentences and generate more accurate coordinated tags.

Module 2: Feature Selection and Feature Extraction

Feature selection and feature extraction are important steps in preparing the input data for a structural time-aware coordinated tag generation based on transformer network. Here are some ways to approach these tasks, Feature Selection: In feature selection, you identify the most relevant features from the input data that will be used by the model. For structural time- aware coordinated tag generation, some possible features to consider include, Word embeddings These are representations of words as vectors in a high-dimensional space. Word embeddings capture the semantic meaning of the words, which can be useful for coordinated tag generation. Part-of-speech tags These tags indicate the grammatical role of each word in the sentence, which can help the model to identify relationships between words. Dependency relations These relations describe the syntactic structure of the sentence and can be used to identify coordinated phrases. Sentence position the position of each word in the sentence can be a useful feature for coordinated tag generation, as it can help the model to identify the temporal structure of the sentence. Feature Extraction: In feature extraction, you create new features from the existing input data that may be more relevant or useful for the model. Some possible techniques for feature extraction include Dimensionality reduction This involves reducing the number of features in the input data by projecting it onto a lower- dimensional space. This can help to eliminate noise and improve the efficiency of the model. Text normalization: This involves converting text to a standard format, such as lowercasing all letters, removing punctuation and stop words, and stemming or lemmatizing words. This can help to reduce the sparsity of the input data and improve the performance of the model. Contextual embeddings: These are embeddings that take into account the context of the words in the sentence. They can be generated using pre-trained models such as BERT or ELMo. Overall, feature selection and extraction are important for improving the performance of a structural time-aware coordinated tag generation based on transformer network model. The choice of features will depend on the specific task and the available data.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

Module 3: Model Training, Testing and Prediction

Model training, testing and prediction are important steps in developing a structural time-aware coordinated tag generation based on transformer network. Here is an overview of these steps:

Model Training and Testing:

Define the model architecture: The first step is to define the architecture of the transformer-based model that will be used for coordinated tag generation. This involves specifying the number and type of layers, the size of the hidden layers, and other hyperparameters Prepare the training data: The next step is to prepare the input data for training the model, which includes feature selection, feature extraction, and data splitting. This data is used to train the model to predict the coordinated tags for a given sentence. Train the model: The model is trained using the prepared data. During training, the weights of the model are adjusted to minimize the difference between the predicted coordinated tags and the actual tags. Validate the model: After training, the model should be validated using a separate validation dataset to ensure that it is not overfitting to the training data. If the validation performance is not satisfactory, the model may need to be retrained with different hyperparameters or data preparation techniques.

Model Prediction:

Prepare the test data: The test data is prepared in the same way as the training data, but it is not used during the training phase. The test data is used to evaluate the performance of the model on unseen data. Predict coordinated tags: The model is used to predict the coordinated tags for each sentence in the test data. The predicted tags are compared with the actual tags to evaluate the accuracy of the model. Evaluate the model: The performance of the model is evaluated using standard metrics such as precision, recall, and F1 score. These metrics help to determine the effectiveness of the model and identify areas for improvement. Overall, model training and prediction are iterative processes that involve tuning the model architecture and hyperparameters to achieve the best performance. It is important to carefully monitor the performance of the model during both the training and prediction phases to ensure that it is providing accurate and reliable coordinated tags for natural language sentences.

Figure: Testing process

V. CONCLUSION

In conclusion, a structural time-aware coordinated tag generation based on transformer network is a powerful approach for natural language processing that takes into account the temporal structure of the input data. This type of model can be used to accurately predict coordinated tags for sentences in a wide range of applications, from machine translation to sentiment analysis. To develop a successful model, it is important to carefully prepare the data, select relevant features, and train and validate the model using appropriate techniques. The transformer-based architecture provides several advantages over traditional approaches, including the ability to process long sequences of text and the use of attention

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 8, April 2023

mechanisms to selectively focus on important parts of the input. Overall, a structural time-aware coordinated tag generation based on transformer network is a promising approach for natural language processing that has the potential to significantly improve the accuracy and efficiency of many applications. As research in this area continues to advance, we can expect to see even more sophisticated models that can handle increasingly complex and nuanced language tasks.

REFERENCES

[1]. Y. Mehmood and V. Balakrishnan, An enhanced lexicon-based approach for sentiment analysis: A case study on illegal immigration, Online Inf. 0295.

[2]. Y. Oksuz and E. Demir, Comparison of open-ended questions and mul- tiple choice tests in terms of psychometric features and student perfor- 10.16986/HUJE.2018040550.

[3]. L. Galhardi, H. Senefonte, [3] D. S. Thom, and J. R. Brancher, Exploring distinct features for automatic short answer grading, in Proc. Conf., doi: 10.5753/eniac.2018.4399.

[4]. M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, SemEval-2014 task 1: Evaluation of compositional dis- tributional semantic models on full sentences through semantic related- ness and textual entailment, in Proc. 8th Int. Workshop Semantic Eval.

[5]. S. Jordan and T. Mitchell, E-assessment for learning? The potential of short-answer free- text questions with tailored feedback, Brit. J. Educ.

[6]. O. Bukai, R. Pokorny, and J. Haynes, An automated short-free-text scor- ing system: Development and assessment, in Proc. 20th Interservice/Ind.

[7]. N. Othman, R. Faiz, and K. Smali, Manhattan siamese LSTM for question retrieval in community question answering, in Proc. Int. Conf. [30] U. Masaki and U. Yuto, Automated short-answer grading using deep neural networks and item response theory, in Proc. Int. Conf. Artif. Intell.

[8]. C. Sung, T. Dhamecha, S. Saha, T. Ma, V. Reddy, and R. Arora, Pre- training BERT on domain resources for short answer grading, in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. 10.18653/v1/D19- 1628.

[9]. J. Pennington, R. Socher, and C. Manning, Glove: Global vectors for word representation, in Proc. Conf. Empirical Methods Natural Lang.