A Natural Language Processing Model for the Development of an Italian-Language Chatbot for Public Administration

Antonio Piizzi[1], Donatello Vavallo[2], Gaetano Lazzo[3], Saverio Dimola[4], Elvira Zazzera[5]
Tempo S. R. L., Bari, Italy[1, 2, 3, 4]
Kad3 S. R. L., Fasano, Italy[5]

***Abstract*—Natural Language Processing models (NLP) are used in chatbots to understand user input, interpret its meaning, and generate conversational responses to provide immediate and consistent assistance. This reduces problem-solving time and staff workload and increases user satisfaction. There are both rule-based chatbots, which use decision trees and are programmed to answer specific questions, and self-learning chatbots, which can handle more complex conversations through continuous learning about data and user interactions. However, only a few chatbots have been developed specifically for the Italian language. *T*he development of chatbots for Public Administration (PA) in the Italian language presents unique challenges, particularly in creating models that can accurately understand and respond to user queries based on complex, context-specific documents. This paper proposes a novel natural language processing (NLP) model tailored to the Italian language, designed to support the development of an advanced Question Answering (QA) chatbot for PA. The core of the proposed model is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, enhanced with an encoder/decoder module and a highway network module to improve the filtering and processing of input text. The principal aim of this research is to address the gap in Italian-language NLP models by providing a robust solution capable of handling the intricacies of the Italian language within the context of PA. The model is trained and evaluated using the Italian version of the Stanford Question Answering Dataset (SQuAD-IT). Experimental results demonstrate that the proposed model outperforms existing models such as BIDAF in terms of F1-score and Exact Match (EM), indicating its superior ability to provide precise and accurate answers. The comparative analysis highlights a significant performance improvement, with the proposed model achieving an F1-score of 59.41% and an EM of 46.24%, compared to 49.35% and 38.43%, respectively, for BIDAF. The findings suggest that the proposed model offers substantial benefits in terms of accuracy and efficiency for PA applications.**

Keywords—Natural Language Processing; chatbot; BERT; transformer; Italian language

I. Introduction

A chatbot is an application that uses various techniques to understand user input, interpret its meaning, and generate responses in a conversational manner based on the user input. Implementing a chatbot as a support tool offers several advantages for both the end users and the businesses [1]. First, chatbots can provide immediate assistance to users by helping to reduce the time it takes to solve problems. Second, chatbots can provide 24-hour support, even outside normal business hours, with shorter resolution times and higher user satisfaction. In addition, chatbots can be used to identify common problems and trends through analysis of user questions, enabling the company to proactively address these issues and reduce the number of support requests. It is also cost-effective as it helps reduce the workload of IT staff and the need for additional staff or overtime. A chatbot also offers benefits to the business because it helps reduce the workload of IT staff and the need for additional staff or overtime. Chatbots can be classified into two main variants: rule-based and self-learning. Rule-based chatbots are developed to answer specific questions or perform actions based on predefined rules and logic because the responses are written in advance and correspond to a set of predefined questions or commands. However, these chatbots are limited to understanding complex natural languages and are unable to adapt to new situations. Therefore, chatbots based on Artificial Intelligence (self-learning chatbots) have been developed to overcome these issues. Self-learning chatbots learn from data and user interactions, gradually improving their ability to respond accurately. They can manage more complex conversations and adapt to new scenarios and user requests. However, they require initial training and are more complex to develop than rule-based chatbots. Complexity is to ensure the best response to each and every unpredictable user input. In the context of public administration (PA), a self-learning chatbot should be capable of understanding and interpreting documents provided by users as input in order to generate appropriate responses. For example, such a chatbot could handle requests that involve reading and interpreting official documents, offering precise and relevant answers. A chatbot developed with these requirements would represent a great advantage in terms of time as well as efficiency for public administration staff.

Despite the widespread use of chatbots in various languages, there is a significant gap in the development of advanced NLP models specifically designed for the Italian language, particularly in the context of Public Administration (PA). Most existing chatbots are either rule-based, which limits their ability to handle complex and varied user inputs, or they are designed for languages like English, for which extensive datasets and pre-trained models are available. However, Italian-language chatbots remain underdeveloped due to the lack of comprehensive datasets and specialized models that can accurately process and understand Italian text, especially within the specific and formal context of PA.

This research seeks to address this gap by proposing a novel NLP model based on the BERT architecture, tailored specifically for the Italian language. The model is designed to overcome the limitations of existing approaches by incorporating an encoder/decoder module and a highway network module to enhance the processing of Italian text. By training and evaluating the model on the SQuAD-IT dataset, this study aims to provide a robust solution that significantly improves the accuracy and efficiency of Italian-language chatbots in PA contexts, thereby filling a critical void in current NLP research. In particular, the proposed work presents an NLP model for the development of a QA self-learning chatbot able to read any document and provide relevant and specific responses to users. The proposed model architecture is based on the BERT [2] model architecture, with the addition of an encoder/decoder module and a highway network module to improve the filtering of the input text. Specifically, the encoder module takes as input a sequence of tokens and transforms them into a dense vector representation that captures the semantic and syntactic information of the text. The decoder module takes the vector representation provided by the encoder as input and generates a new sequence of text. The highway network filters irrelevant information before processing the last dense layers. Moreover, the proposed model is trained on the SQuAD-IT dataset to develop a chatbot specifically for the Italian language. Only a few works have used the Italian version of the SQuAD dataset, and the proposed work is intended to present an efficient and suitable architecture for developing an Italian-specific chatbot. The main contributions of the proposed work are summarized below.

1) A customized BERT model architecture with the addition of an encoder/decoder module and a highway network module has been proposed. The proposed model is developed to be integrated into an Italian-specific chatbot to improve the work of PA staff by reducing time and errors in processing and understanding documents.

2) The proposed model is trained and tested on the Italian version of the SQuAD dataset to evaluate the model's ability to process the Italian language.

3) The results of the experiments conducted on the SQuAD-IT dataset show that the proposed model has a good ability to provide exactly the expected answers. Moreover, a comparative analysis shows that the proposed model outperforms compared to other NLP models, such as BIDAF.

II. Related Work

Different types of self-learning chatbots have been developed based on the deep-learning models used. Three macro-categories can be identified: chatbots based on Convolutional Neural Networks (CNNs), chatbots based on Recurrent Neural Networks (RNNs), and chatbots based on hybrid models. CNNs are mainly used for pattern recognition in text data, such as sentences or paragraphs. Subsequently, sentence similarities between question-answer pairs are used to assess relevance and rank all candidate answers. In study [3] a CNN for learning an optimal representation of question-and-answer sentences has been proposed. The CNN encodes the correspondences between words to better acquire interactions between questions and answers, resulting in a significant increase in accuracy. The proposed CNN consists of two distributional sentence models based on convolutional neural networks (ConvNets) that map question-and-answer sentences into their distributional vectors, which are then used to learn their semantic similarity. In study [4], a model that considers both similarities and differences between question-and-answer by decomposing and composing lexical semantics on sentences has been proposed. Given a pair of sentences, the model represents each word as a low-dimensionality vector and computes a semantic correspondence vector for each word based on all the words in the other sentence. Then, based on the semantic correspondence vector, each word vector is decomposed into two components: similar and dissimilar. The similar components of all words are used to represent the similar parts of the sentence pair and the dissimilar components of each word are used to model the dissimilar parts explicitly. Subsequently, the similar and dissimilar components are composed into a single feature vector that is used to predict sentence similarity. In study [5], a CNN with a Siamese structure with two sub-networks processing the question and the candidate response has been proposed. The input is a sequence of words each of which is translated into its corresponding distributional vector, producing a matrix of sentences. Convolutional feature maps are applied to this matrix of sentences, followed by ReLU activation and simple max-pooling to achieve a representation vector for the query and candidate response. CNNs can have difficulty capturing long-term dependencies in text. Therefore, they are combined with RNNs or transformers to improve performance.

RNNs can process data sequences of variable length, making them ideal for text. In research [6], a bilateral multi-perspective matching (BiMPM) model has been proposed. The proposed model encodes the sentences through a bidirectional Long Short-Term memory Network (BiLSTM). The matching between encoded sentences is aggregated through a layer BiLSTM into a matching vector of fixed length. It is used in the last fully connected layer of the proposed model to make a decision. The approach in study [7] extends a long short-term memory (LSTM) proposed as a holographic dual LSTM (HDLSTM). HDLSTM is a unified architecture for both deep sentence modeling and semantic matching. RNNs are less efficient than CNNs and transformers in terms of parallelization and training speed. Therefore, hybrid models that combine the advantages of each model have been developed. RNNs have been combined with CNN [8], with attention mechanisms [9] or with transformer [10], [11]. Minjoon Seo et al. [12] have proposed a Bi-Directional Attention Flow network (BIDAF) that uses a bi-directional attention mechanism to obtain a query-sensitive context representation. The attention layer is not used to summarize the context paragraph into a vector of fixed size, but attention is computed for each time step, and the expected vector in each time step, together with the representations of the previous layers, can flow through the next modeling layer.

The introduction of the transformers in NLP models has led to advantages in efficiency, ability to capture long-range relationships, parallelization, and quality of representations. The transformer architecture allows higher parallelization during training and inference because it does not require sequential computation as in RNNs. This significantly reduces training time compared with RNNs. Therefore, using transformers represents one of the best choices to develop self-learning chatbots.

The proposed model is based on BERT architecture that use transformer to construct deep and bidirectional representations of words so as to capture complex contextual relationships. Moreover, the proposed model implements additional modules in the standard BERT architecture to increase the model's capabilities in providing correct answers.

III. Materials and Methods

The proposed model is based on an extended version of the BERT model by adding an encoder/decoder module and a highway network module. The proposed model is illustrated in Fig. 1. The proposed model consists of four main modules.

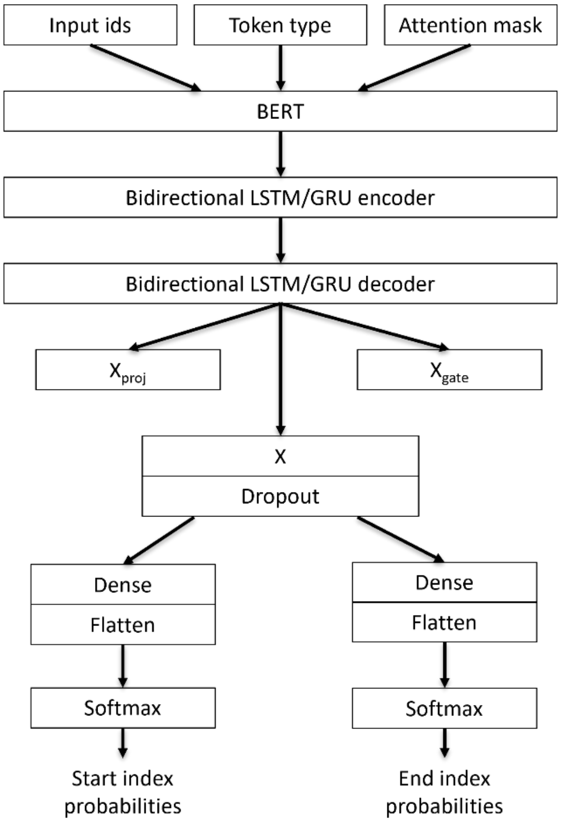

Fig. 1. Proposed model architecture.

A. Input Module

The first module consists of BERT architecture to encode the input into a vector representation that is then processed by the subsequent structures. BERT architecture can be represented as a multilayer bidirectional transformer encoder. BERT's pre-training is based on two different unsupervised tasks: the Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, a portion of the words in the text are masked, and the model must predict them. In NSP, the model must determine whether two sentences appear consecutive in the original text. This pre-training makes the BERT model scalable (fine-tuned) for different tasks, such as QA. BERT takes as input the combinations of the question and context as a single embedded sequence. The input embeddings are the sum of the token embeddings and segment embeddings. Specifically, token embeddings represent the encoding of the question into an embedding vector, and segment embeddings represent vectors indicating the segment to which each token corresponds. The segment embeddings are used to distinguish between question and context in the input text. Let $x = [x_1, x_2, \ldots, x_n]$ represents a sequence of input, and e_i represents an embedded achieved combining the token and segment embeddings for each x_i . The sequences of embeddings $E = [e1, e_2, \ldots, e_n]$ is the input of the BERT module. The module BERT processes the embedding sequence through L transformer layers to obtain the output sequences $H^L = [h_1^L, h_2^L, \ldots, h_n^L]$ where h_i^L is the hidden representation of $x_1$at the L level.

B. Encoder / Decoder Module

The encoder/decoder module consists of two sequential BiLSTM layers. The introduction of this module better captures the context and temporal sequence of words, thus improving the model's overall performance in understanding. Specifically, the BERT output $H^L = [h_1^L, h_2^L, \ldots, h_n^L]$ is taken as input to encoder/decoder module to produce a new sequence of hidden representations $H^{BiLSTM} = \left[h_1^{BiLSTM}, h_2^{BiLSTM}, \ldots, h_n^{BiLSTM}\right]$.

C. Filter Module

The highway network module aims to filter out irrelevant information before processing the last dense layers. Highway Network transformations are based on a linear combination between the non-linear transformation of the input and the original input following a gate function. The output of the Highway network module is defined in Eq. (1), where ∘ represents the element-by-element multiplication.

$$y_i = T\left(h_i^{BiLSTM}\right) \circ S\left(h_i^{BiLSTM}\right) + \left(1 - T\left(h_i^{BiLSTM}\right)\right) \circ h_i^{BiLSTM} \quad (1)$$

The linear transformation $T(\cdot)$ is defined by Eq. (2), where W_T and b_T are the weights and the bias of the gate function, respectively, and σ represents the sigmoid function.

$$T\left(h_i^{BiLSTM}\right) = \sigma\left(W_T h_i^{BiLSTM} + b_T\right) \quad (2)$$

The non-linear transformation $S(\cdot)$ is defined in Eq. (3), where W_S and b_S are the weights and the bias of the non-linear transformation, respectively, and $ReLU$ is the activation function.

$$S\left(h_i^{BiLSTM}\right) = ReLU\left(W_S h_i^{BiLSTM} + b_S\right) \quad (3)$$

D. Output Module

The output module consists of two fully connected layers with softmax activation function. The output module predicts the start and end positions of the response within the context following Eq. (4) and Eq. (5), respectively.

$$P_{start}(i) = softmax(W_{start}y_i + \ b_{start}) \qquad (4)$$

where, W_{start} and b_{start} represent the weights and bias of the fully connected layer for the prediction of the start token.

$$P_{end}(i) = softmax(W_{end}y_i + \ b_{end}) \qquad (5)$$

where, W_{end} and b_{end} represent the weights and bias of the fully connected layer for the prediction of the end token.

IV. Dataset

SQuAD-IT [13] dataset is a translated and adapted Italian version of the popular SQuAD dataset [14] SQuAD dataset has been developed to evaluate NLP models in English. The main advantage of this dataset is that it is realistic because humans crowdsourced it manually. It includes 536 English Wikipedia articles with more than 100,000 related question-answer pairs. Each crowd-worker was asked to answer up to five questions about a Wikipedia passage, highlighting the answer in the passage. Each question is associated with a segment of text within the article (context), and the answer is a subset of the context. Each question had several specific answers provided by different people. In SQuAD-IT, the texts, questions, and answers in SQuAD have been translated into Italian. The translation aims to maintain the same structure and content as the original dataset but ensures that the sentences are natural and grammatically correct in Italian. SQuAD-IT has been developed to evaluate NLP models that are not limited to English. The use of the SQuAD-IT dataset for the experiments enables the evaluation of the proposed model in understanding the Italian language to develop an Italian-specific chatbot.

V. Experimental Evaluation

Experiments have been conducted to evaluate the proposed model's performance in providing the correct answer in the QA task. The performance has been evaluated in terms of F1-score and Exact Match (EM). F1-score is the harmonic mean of accuracy and recall rate. In other words, the F1-score measures the overlap between the words in the predicted answer and the corresponding words in the correct answer (ground truth). The EM computes the percentage of correct answers generated by the model compared to the ground truth. F1-score and EM have been computed following Eq. (6) and Eq. (7), respectively.

$$F1 - score = \ 2 \cdot \frac{precision \cdot recall}{precision + recall} \qquad (6)$$

$$EM = \frac{\sum_{i=1}^{N} I(pred_i = truth_i)}{N} \qquad (7)$$

In Eq. (7), N represents the number of test examples, $pred_i$ and $truth_i$ represent the predicted answer and the correct answer, respectively, and $I(\cdot)$ is a function that returns 1 if the condition is truth or 0 otherwise.

To evaluate the proposed model, the SQuAD-IT dataset has been divided following the train-test split ratio of 80:20, and the proposed model has been trained following the parameters detailed in Table I. The results of the proposed model in terms of F1-score and EM are shown in Table II. The proposed model achieves good performance in providing exactly the correct answer, even considering that it is obtained on an Italian dataset explored by very few studies in the literature. The F1-score of 59.41% indicates a good balance between precision and recall, and an EM of 46.25% indicates that almost half of the answers provided by the model are perfectly accurate. An EM score lower than the F1-score means that the model often approximates correct answers but does not provide correct answers. EM is a very rigorous metric because it measures the percentage of answers that are exactly the same as the correct answers.

TABLE I. Training Parameters for the Proposed Bert-Based Model

Parameter	Value
Encoding dimension	128
Decoding dimension	64
Loss	Sparse Categorical Cross-Entropy
Optimizer	Adam
Batch size	8
Learning rate	5e-5
Number of epochs	6
Dropout	False

TABLE II. Performance Results of the Proposed Bert-Based Model on the Squad-It Dataset

Metric	Score (%)
F1-score	59.4106
EM	46.2544

Moreover, a comparative analysis has been conducted to evaluate the performance of the proposed model compared to the BIDAF model. To a fair comparison, the BIDAF model is trained and tested with the same train-test split ratio used for the proposed model. Table III details the training model parameters used for BIDAF. The comparative analysis shows that the proposed model achieves an improvement of 10.06% in F1-score and 7.81% in EM compared to BIDAF, as shown in Table IV. In other words, the proposed model is capable of providing significantly more correct answers compared to the BIDAF model.

TABLE III. Training Bidaf Model Parameters

Parameter	Value
Loss	Sparse Categorical Cross-Entropy
Optimizer	Adam
Batch size	10
Learning rate	5e-4
Number of epochs	10
Dropout	0.2

TABLE IV. Comparative Analysis

Model	F1-score	EM
Proposed	**59.4106**	**46.2544**
BIDAF	49.3504	38.4313

VI. Discussion

The development of an Italian-language chatbot tailored for Public Administration represents a significant step forward in the application of NLP models to non-English languages. Throughout this study, it became clear that the unique linguistic and contextual challenges of the Italian language, especially in formal and legal settings, demand specialized models that go beyond generic NLP solutions. Our personal insight is that the integration of an encoder/decoder module and a highway network module into the BERT architecture not only enhances the model's ability to process complex input but also addresses the specific needs of the Italian language, which is often syntactically richer and more flexible than English. This study has demonstrated the potential of these modifications to improve the accuracy and relevance of responses in a chatbot context, which is crucial for the efficiency of Public Administration tasks.

Looking towards the future, several areas offer promising opportunities for further exploration. Firstly, the proposed model could benefit from ensembling strategies, where multiple models are combined to refine the accuracy and robustness of the responses. This could help mitigate the limitations observed in exact match accuracy, especially in cases where the model approximates but does not fully capture the correct answers. Additionally, expanding the dataset beyond the SQuAD-IT to include domain-specific data from Public Administration could enhance the model's contextual understanding and make it even more effective in real-world applications. Another avenue worth exploring is the application of transfer learning techniques, where the model could be pre-trained on broader Italian-language datasets and then fine-tuned on Public Administration-specific data. This approach could further improve the model's performance in specialized contexts. Finally, considering the rapid advancements in NLP, future work could also involve adapting the proposed model to more recent architectures like GPT, which might offer even greater capabilities in terms of natural language understanding and generation.

VII. Conclusion

An extended architecture of the BERT model specific to understanding the Italian language has been proposed. The proposed model introduced an encoder/decoder module and a Highway network module before the fully-connected layers into the BERT architecture to improve the ability of the model to capture the context and temporal sequence and to filter out irrelevant information. The proposed model performs well in providing the exact expected answers. Comparative analysis shows that the proposed model outperforms the BIDAF model, providing almost 8% more correct answers. Moreover, the proposed model is one of the first models developed specifically to be able to process Italian language texts, as it was tested with the Italian version of the popular SQuAD dataset. The proposed model represents the ground for developing an Italian-specific chatbot for the PAs. Therefore, the proposed model, implemented in a chatbot, could make the work of PAs more efficient in terms of time and workload reduction. In future, ensembling strategies could be used by combining responses from multiple models to improve overall accuracy and the number of incorrect answers.

Acknowledgment

This research was funded by Regione Puglia - Progetto NOVA PA - Contabilità ed E-procurement per una P.A. Digitale - POR PUGLIA FESR 2014-2020 - Grant number: SJA8705.

References

[1] M. Pislaru, C. S. Vlad, L. Ivascu, e I. I. Mircea, «Citizen-Centric Governance: Enhancing Citizen Engagement through Artificial Intelligence Tools», Sustainability, vol. 16, fasc. 7, p. 2686, 2024.

[2] J. Devlin, M.-W. Chang, K. Lee, e K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». arXiv, 2018.

[3] A. Severyn e A. Moschitti, «Modeling Relational Information in Question-Answer Pairs with Convolutional Neural Networks». arXiv, 2016.

[4] Z. Wang, H. Mi, e A. Ittycheriah, «Sentence Similarity Learning by Lexical Decomposition and Composition». arXiv, 2016.

[5] R. Sequiera et al., «Exploring the Effectiveness of Convolutional Neural Networks for Answer Selection in End-to-End Question Answering». arXiv, 2017.

[6] Z. Wang, W. Hamza, e R. Florian, «Bilateral Multi-Perspective Matching for Natural Language Sentences». arXiv, 2017.

[7] Y. Tay, M. C. Phan, L. A. Tuan, e S. C. Hui, «Learning to Rank Question Answer Pairs with Holographic Dual LSTM Architecture», in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku Tokyo Japan, 2017, pp. 695–704.

[8] M. M. A. Zaman e S. Z. Mishu, «Convolutional recurrent neural network for question answering», in 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), Khulna, 2017, pp. 1–6.

[9] W. Wang, N. Yang, F. Wei, B. Chang, e M. Zhou, «Gated Self-Matching Networks for Reading Comprehension and Question Answering», in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, 2017, pp. 189–198.

[10] T. Shao, Y. Guo, H. Chen, e Z. Hao, «Transformer-Based Neural Network for Answer Selection in Question Answering», IEEE Access, vol. 7, pp. 26146–26156, 2019.

[11] M. Kamyab, G. Liu, e M. Adjeisah, «Attention-Based CNN and Bi-LSTM Model Based on TF-IDF and GloVe Word Embedding for Sentiment Analysis», Appl. Sci., vol. 11, fasc. 23, p. 11255, nov. 2021.

[12] M. Seo, A. Kembhavi, A. Farhadi, e H. Hajishirzi, «Bidirectional Attention Flow for Machine Comprehension». arXiv, 2016.

[13] D. Croce, A. Zelenanska, e R. Basili, «Neural Learning for Question Answering in Italian», in AI*IA 2018 – Advances in Artificial Intelligence, vol. 11298, C. Ghidini, B. Magnini, A. Passerini, e P. Traverso, A c. di Cham: Springer International Publishing, 2018, pp. 389–402.

[14] P. Rajpurkar, J. Zhang, K. Lopyrev, e P. Liang, «SQuAD: 100,000+ Questions for Machine Comprehension of Text». arXiv, 2016.