Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Review of the Development of Input Word Prediction

Qianwei Ke

School of computer science and engineering Central South University Changsha, China

8208210828@csu.edu.cn

Abstract. As one of the important applications in the field of human-computer interaction, input word prediction has made great progress in recent years.This paper reviews and summarizes the development process of input word prediction technology, from the early statistical model-based method to the current deep learning-based technology application, describes its development trajectory and the representative algorithms of each technology.In addition, this paper also analyzes the current problems and challenges faced by input method word prediction, such as language model modeling, user personalized needs, multi-language input and other aspects of the problem, and discusses the future development trend, including the combination of multi-modal information, fusion reinforcement learning and other new technologies.Finally, this paper also looks forward to the extensive application prospects of input word prediction technology in many fields, and its potential contribution to improving user input efficiency, improving user experience and promoting the development of natural language technology.

Keywords: input method; forecast; natural language; processing technology; artificial intelligence; develop.

1. Introduction

Word prediction is a kind of human-computer interaction technology, as a branch of the field of natural language processing (NLP), which aims to predict the word or phrase that the user may enter next by analyzing the text content and context information, so as to improve the input efficiency and accuracy.This technology helps users to enter more quickly by automatically suggesting possible word choices [1].The word prediction technology of input method has developed from manual rules and dictionaries to statiscy-based models (such as N-gram model, etc.), and then the application of deep learning. These include techniques based on machine learning (support vector machines (SVM), naive Bayes classifiers, random forests, etc.), neural network methods (Transformer model and its variants such as Bidirectional Encoder Representations from Transformers (BERT), Generative Pretrained Transformer (GPT), etc.), and now personalization and multimodal input. The development of input word prediction technology has continuously improved the accuracy, intelligence level and user experience of prediction [2,3]. The development of each stage is shown in Figure 1. The development of input method language prediction technology can improve the user's text input efficiency, promote information transfer and knowledge sharing, and promote the development of natural language processing technology.This paper reviews and summarizes the development of input word prediction technology. In addition, this paper also looks forward to the extensive application prospects of input word prediction technology in many fields, and its potential contribution to improving user input efficiency, improving user experience and promoting the development of natural language technology.

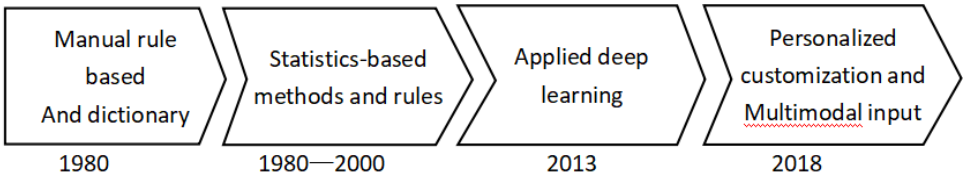

Figure 1. History of Input Prediction Technology Development (Photo/Picture credit : Original)

2. Development History

2.1. Statistics-based methods and rules

In the early 1880s, the early word prediction techniques of input methods were mainly based on manual rules and dictionaries. However, due to the complexity and diversity of languages [2], creating and maintaining rules is very difficult and consumes a lot of time.

From 1980 to 2000, traditional input word prediction was based on basic statistical methods and rules. As an important technical means in the input method system, word prediction technology based on statistical methods and rules is not only widely used, but also plays an important positive role in improving user input efficiency and reducing input errors [2]. The principle of this method is to realize intelligent prediction of user input behavior and context through comprehensive application of statistical methods and rule processing.First of all, through statistical analysis of the user's historical input data, the system will understand the user's input habits, common words and phrase combinations and other information, and then build a model. This model can use statistical information to determine the frequency of certain words in a particular context, and thus predict the words that the user is likely to type. Secondly, you can define some grammatical rules or semantic rules artificially. The system uses rules to process the first few characters or words currently entered, and uses language rules and context information to infer the next word that the user may enter.

The input method word prediction technology based on statistical methods and rules can apply a variety of representative model methods, such as N-gram model based on the statistical frequency of word sequences, hidden Markov model (HMM) for modeling time series data, and so on. Among them, N-gram model is simple, easy to implement and has high computational efficiency, which is suitable for short text prediction and small-scale corpus. However, the N-gram model cannot capture long distance dependencies, requires a large amount of data to accurately estimate parameters, and has limited ability to deal with rare words and unknown contexts. The HMM can handle the data with time sequence characteristics, and can be used to model the complex sequence generation process, and has a good effect on the sequence prediction of finite state space. But similarly, the HMM is difficult to estimate the model parameters, and requires a large amount of training data. At the same time, there may be local optimal problems, and the operation complexity is high.

All in all, the word prediction technology based on statistical methods and rules plays an important role in improving input efficiency and reducing input errors, and has important theoretical and practical significance for the optimization and improvement of input system. However, this method ignores the order and semantic information of words, highly relies on manual design, and can not deal with the polysemy between words and complex linguistic phenomena.

2.2. Applying deep learning

Machine learning: Word prediction technology based on machine learning is to use machine learning algorithms to predict the user's next possible word in the input method. By analyzing information such as the user's input context and historical data, the technology can learn input patterns and word probability distributions to provide intelligent input predictions and recommendations. The principle and flow of this technology are shown in Figure 2.

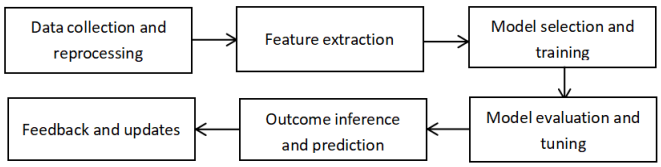

Figure 2. Flowchart of Word Prediction With Machine Learning Input Method (Photo/Picture credit : Original)

In the data collection and preprocessing stage in natural language processing tasks, a large amount of data must first be collected and prepared as training corpus. This includes obtaining pre-training data for language models, domain-specific corpora, and user input text and related contextual information. Subsequently, the collected data is preprocessed, such as word segmentation, stop word removal, and normalization. In the feature extraction stage, useful features need to be extracted from the preprocessed data to effectively represent the input context. These features include n-gram features, part-of-speech tagging features, contextual features, etc., which are then converted into numerical feature vector representations. Common transformation methods include bag-of-words models and TF-IDF (term frequency-inverse document frequency).

In the model selection and training phase, appropriate machine learning models are selected for training based on task requirements, such as SVM, naive Bayes classifiers, random forests, neural networks, etc., and the extracted feature vectors are used for training. During training, the model learns the relationship between inputs and outputs. Then, in the model evaluation and tuning stage, the test set is used to evaluate the trained model. The evaluation indicators can be accuracy, precision, recall, etc. Based on the evaluation results, the model is tuned, such as adjusting hyperparameters, feature selection, etc., to improve prediction accuracy and performance.

In the inference and prediction phase, the trained model is used to calculate the probability distribution of each possible word and provide the user with suggestions for the next possible word. Finally, during the feedback and update phase, the system continuously optimizes the model based on user feedback and input behavior to improve prediction accuracy and user experience. By continuously updating model parameters and training data, the model is made smarter and more in line with user needs.

Neural Network: In recent years, input method word prediction technology based on neural networks has achieved significant development in the field of artificial intelligence. With the continuous development of deep learning technology, more and more research has begun to use deep neural network models, and Transformer models and large-scale pre-training models have emerged, such as BERT, GPT series, etc. [2]. Input method word prediction technology based on neural networks predicts the next word that the user may input by analyzing the user's current input text and contextual information, and provides candidate word recommendations. The main idea of this technology is to treat the word prediction task of the input method as a sequence generation problem, and use a neural network model to learn the contextual relationship between words, so that the user's next input can be predicted more accurately. The process of input method word prediction technology based on neural network is shown in Figure 3.

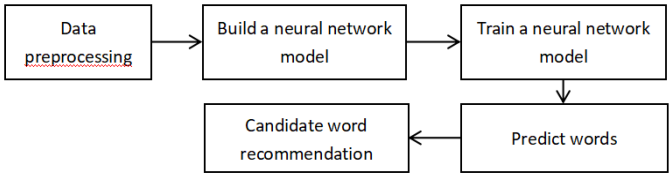

Figure 3. Flowchart of Word Prediction based on Neural Network Input Method (Photo/Picture credit :Original)

To achieve input method word prediction based on neural networks, you first need data preprocessing. You need to collect large-scale text data and perform word segmentation processing to divide the text into words as training samples.Then build a neural network model and select an appropriate neural network model, such as RNN or variants (such as LSTM, GRU) [4], Transformer or its variants BERT model, GPT series models, etc., are used to learn contextual relationships between words.Then the preprocessed training data is input into the neural network model for training, and the parameters of the neural network are continuously adjusted through the back propagation algorithm so that it can better predict the next possible word. In actual use, the input method will extract the most recent text as context based on the text that the user has entered, and input the context into the trained neural network model.The model will calculate the possible next word based on the context information and

the learned word relationships, and give a probability distribution. According to the probability distribution output by the model, several words with higher probability are selected as candidate words to recommend to the user. Some other factors, such as word frequency, user input habits, etc., are usually also considered to sort and filter candidate words to provide more accurate prediction and recommendation results.

The Transformer model is a deep learning model for processing sequence data, which is particularly suitable for processing natural language processing tasks [5, 6]. The Transformer model introduces a self-attention mechanism to better capture long-distance dependencies in sequences, abandoning the recursive structure of traditional sequence models such as RNN and LSTM, allowing the model to process input sequences in parallel, significantly It improves computational efficiency and has better results when processing long sequences [5]. In addition, the hierarchical structure of the Transformer model includes multi-layer encoders and decoders. Each layer is composed of a multi-head attention mechanism and a feed-forward neural network, allowing the model to extract and combine input information layer by layer, allowing it to learn more complex representation of characteristics. This is very important for many word prediction tasks such as machine translation, etc. Therefore, the Transformer model has achieved great success in the field of natural language processing and has become one of the important milestones in modern deep learning. Transformer model architecture is shown in Figure 4.

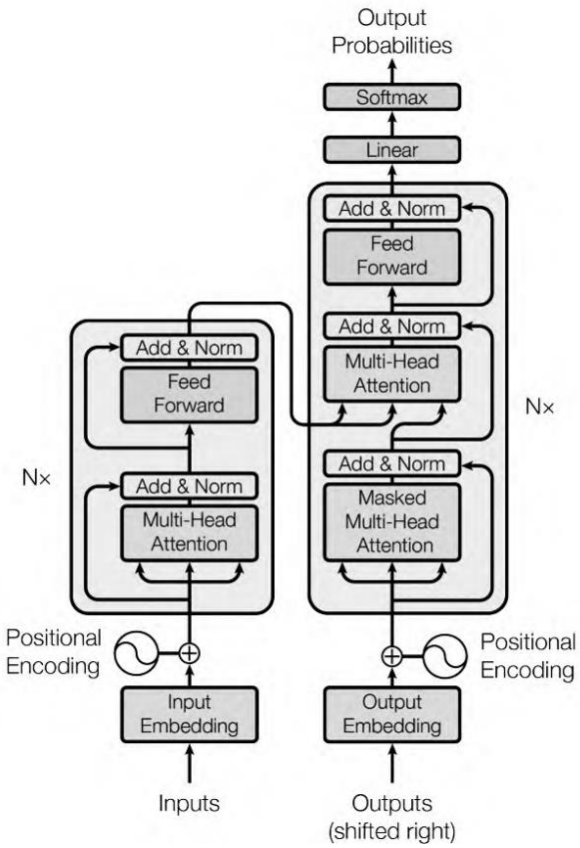

Figure 4. Architecture of The Transformer Model [7]

The BERT model is a pre-trained language model based on the Transformer architecture. The core innovation of the BERT model is the introduction of a bidirectional training method, which can simultaneously consider left and right context information to better capture the association between words.Compared with traditional context-free language models, BERT can better understand the context and meaning of words in sentences, helping to improve the performance of natural language processing tasks. The BERT model achieves good results by pre-training on a large-scale text corpus, performing self-supervised learning based on massive text data, learning rich word and sentence representations, and then fine-tuning on specific tasks.Through this pre-training-fine-tuning method,

various downstream tasks can be transferred and learned, such as text classification, named entity recognition, etc. It has achieved breakthrough results in multiple NLP benchmark tests and has become the focus of NLP research [8].

The GPT series of models is also a pre-training model based on the Transformer architecture.Different from BERT, it uses Transformer decoder for one-way (left to right) pre-training. This method of training based on autoregressive (autoregressive) method predicts the next word given the previous text content. This approach makes the model more suitable for generative tasks, such as dialogue generation, article summarization, etc. Currently, GPT-4 is the largest GPT model. Compared with previous versions of the model, it contains more parameters, can generate more realistic text, and has stronger academic capabilities. In addition, GPT-4 shows great advantages in English in the MMLU (Multi-Task Language Understanding) benchmark test, and also shows strong performance in other languages. However, GPT-4 still has major limitations, and it can still produce hallucinations and reasoning errors. Both BERT and GPT series models have achieved great success in the field of natural language processing and have become one of the most advanced pre-trained language models. Each has performed well on different types of tasks and is of great significance to the development of input method word prediction technology.

2.3. Personalized customization and multi-modal input

Since 2018, personalized customization and multi-modal input technology have become increasingly popular in the input method field.Input method prediction technology for personalized customization and multi-modal input refers to using the user's personalized input habits and multiple input methods (such as voice, handwriting, pinyin, etc.) [9] to improve the prediction accuracy and user experience of the input method . This technology combines knowledge from the fields of machine learning, natural language processing, and computer vision to achieve more intelligent and personalized input predictions by analyzing the user's input behavior and contextual information.

The biggest innovation of this technology mainly lies in the feature extraction and multi-modal fusion stages. This technology collects user input data, habits, and multi-modal input information (such as voice, handwriting, pictures, emoticons, etc.) to build a personalized user model. Afterwards, different features are extracted and relevant labels are added to different types of input data, such as acoustic feature extraction for speech input, word segmentation and vectorization for text input, stroke analysis for handwriting input, etc [10]. In the model training stage, this technology comprehensively uses machine learning algorithms, such as neural networks, support vector machines, etc., to train input method prediction models based on the user's personalized data. Finally, multi-modal fusion will be performed to fuse input information from different modalities together and make input predictions based on the user's input methods and habits. In addition to text information, the prediction results also include pictures and emoticons with relevant tags, voice, etc.

Input method prediction technology for personalized customization and multi-modal input can more accurately predict the user's input intention, improve input method efficiency, and reduce input errors by analyzing the user's multi-modal input habits and contextual information. Secondly, personalized customization can provide customized services according to users' personal preferences and habits, meet users' individual needs, and improve user satisfaction and loyalty. At the same time, by combining multiple input methods such as pictures, voice, and handwriting, users can choose the input method that best suits them, increasing input flexibility and convenience, and improving the overall user experience. The most important thing is that the input method can continuously learn user input behavior and feedback, optimize prediction algorithms, improve accuracy and intelligence levels, and continuously improve user experience, making the input process more intelligent, personalized and efficient.

However, there is a user adaptation period in the application of this technology. Users may need a certain amount of time to adapt to personalized customized input suggestions, and there may be an initial period of discomfort and running-in. In addition, input misjudgment is a problem that cannot

be ignored. Personalized customization may cause the input method to rely too much on user habits, thus ignoring changes in user intentions, leading to input misjudgment. At the same time, the processing of multi-modal compatibility also needs to be improved. Different input methods may have compatibility issues, causing some users to be unable to fully utilize the convenience brought by multi-modal input.

3. Problems and Challenges

Current input method word prediction technology faces four major problems and challenges.

First, there are still challenges in context understanding. Although modern technology has made certain progress, it is still difficult to accurately understand information in complex contexts, resulting in inaccurate prediction results [11].

Secondly, users' personalized needs are also an important challenge, because different users have different usage habits and preferences, and traditional technology still has room for improvement in meeting users' personalized needs. In addition, with the development of globalization, input method word prediction technology needs to better support multi-language input [12], which poses new challenges to input method word prediction technology.

At the same time, handling input errors is also one of the challenges faced by current technology [13]. Users often make typos or spelling errors. How to effectively identify and deal with these input errors also needs further improvement.

Finally, when it comes to user sensitive information [14], there are issues such as leakage of sensitive information, mutual interference and leakage of shared device privacy, and cloud processing security. Input method word prediction technology also needs to be strengthened in terms of privacy and security. These are both problems and opportunities for the development of input method word prediction technology. By solving these problems, the accuracy, intelligence level and user experience of input method word prediction technology will be continuously improved to better meet the needs of users.

4. Conclusion

This article comprehensively reviews the development history of input method word prediction technology, from early rule-based and statistical model-based methods to advanced technologies based on deep learning and the emerging trends of personalized customization and multi-modal input in recent years, and analyzes Each type represents the advantages and disadvantages of algorithms. In addition, this article also organizes and analyzes the challenges faced by current technologies, including the complexity of context understanding, the diversity of user personalized needs, support for multi-language input, handling of input errors, and the protection of user privacy and data security. These are not only obstacles to current technological development, but also areas that need to be focused on in the future. In the future, input method word prediction will continue to improve in the direction of further enhancing context understanding capabilities, meeting user personalized needs, protecting user privacy, and cross-modal interaction .

With the continuous development of technology, more innovative technologies and methods will appear in the future. The development of input method prediction technology will further push the boundaries of natural language processing technology and make greater contributions to the development of artificial intelligence.

References

[1] X. Yang, X. Sun, & Z. Dong. A Survey of Word Prediction Techniques for Chinese Input Method. ACM Computing Surveys, 51(5), 1-32,2018.

[2] F. Wu, C. Yang, X. Lan, et al. A Review and Prospect of Artificial Intelligence. China Science Foundation, 32(03), 243-250, 2018.

[3] L. Hao, Z. Yu. Development and Application of Natural Language Processing Technology Based on Artificial Intelligence. Heilongjiang Science, 14(22), 124-126, 2020.

[4] A. Graves, A. Mohamed, & Geoffrey Hinton. Speech recognition with deep recurrent neural networks. https://arxiv.org/pdf/1303.5778.pdf. Accessed December 12, 2023.

[5] L. Lin. Development of Natural Language Processing Technology in the Era of Artificial Intelligence. Electronic World, (22), 24-25, 2020 .

[6] Z. Yang, Zihang Dai, Yu Yang, et al. . Xlnet: Generalized autoregressive pretraining for language understanding. https://arxiv.org/pdf/1906.08237, 2020

[7] T. Ma, Guoliang Zhang, & Xiaojun Guo. (2024). A Review of Research on Offensive and Defensive Natural Language Processing Based on Deep Learning. China-Arab Science and Technology Forum (Bilingual), 2024(01), 98-102.

[8] J. Devlin, Ming-Wei Chang, Kenton Lee, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[9] K. Onishi, Hiroshi Tanaka, & Satoshi Nakamura. Multimodal Voice Activity Prediction: Turn-taking Events Detection in Expert-Novice Conversation. In Proceedings of the 11th International Conference on Human-Agent Interaction 13-21,2023.

[10] Y. Yang. Application Research of Artificial Intelligence Natural Language Processing in Audio Textbooks. Audio Technology, 46(05), 29-35, 2022.

[11] Z. Dai, J. Callan. Deeper text understanding for IR with contextual neural language modeling. In Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, 985-988, 2019.

[12] M. Samson Lakew, M. Cettolo, M. Federico. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. arXiv preprint arXiv:1806.06957, 2018.

[13] A. Can Kinaci. Spelling Correction using recurrent neural networks and character level n-gram. In 2018 International Conference on Artificial Intelligence and Data Processing (IDAP) 1-4.

[14] Z. Zhang, H. Zhao, R. Wang. Machine reading comprehension: The role of contextualized language models and beyond. arXiv preprint arXiv:2005.06249, 2020.