International Journal of Intelligent Information Technologies
Volume 21 • Issue 1 • January-December 2025

Semantic Annotation Model and Method Based on Internet Open Dataset

Xin Gao
https://orcid.org/0009-0002-3363-140X
State Grid Beijing Electric Power Company, China

Yansong Wang
State Grid Beijing Electric Power Company, China

Fang Wang
State Grid Beijing Electric Power Company, China

Baoqun Zhang
State Grid Beijing Electric Power Company, China

Caie Hu
State Grid Beijing Electric Power Company, China

Jian Wang
State Grid Beijing Electric Power Company, China

Longfei Ma
State Grid Beijing Electric Power Company, China

ABSTRACT

Traditional semantic annotation faces the problem of dataset diversity. Different fields and scenarios need to be specially annotated, and annotation work usually requires a lot of manpower and time investment. To meet these challenges, this paper deeply studies the semantic annotation model and method based on internet open datasets, aiming to improve annotation efficiency and accuracy and promote data resource sharing and utilization. This paper selects Common Crawl dataset to provide sufficient training samples; methods such as removing stop words and deduplication are used to preprocess data to improve data quality; a keyword extraction model based on heuristic rules and text context is constructed. In terms of semantic annotation model, this paper constructs a model based on Bidirectional Long Short-Term Memory (BiLSTM), which can make full use of the part-of-speech information of the corpus context, capture the part-of-speech features of the corpus, and generate semantic tags through supervised learning.

KEYWORDS

Internet Open Dataset, Semantic Annotation Method, Semantic Annotation Model, Topic Coverage

INTRODUCTION

With the rapid development of internet technology, the internet has become the main source of information for people, deeply affecting modern life. However, efficiently using these massive amounts of data has become challenging. Although network data contains great value, it is mostly unstructured or semi-structured, which makes it difficult for computers to interpret directly. Therefore, we urgently need to find ways to tap the value of these data, among which semantic labeling technology is particularly critical. By adding meaning and associated information to the data, semantic labeling enables computers to better understand the data and realize data sharing and intelligent processing. In the era of big data, open datasets are increasing day by day, promoting data sharing and knowledge innovation. However, traditional semantic labeling methods still face challenges, such as diverse formats and different requirements (Choi et al., 2021). Therefore, it is essential to explore a semantic labeling model based on open datasets on the internet, which can not only improve labeling efficiency

DOI: 10.4018/IJIIT.370966

but also promote the wider application of data resources and help innovation and progress in various fields.

Data mining and analysis have grown in importance in the internet age for use in business decision-making, scientific research, and other domains. Large, varied, and complicated internet datasets, however, are frequently challenging for conventional data analysis techniques to handle. Rich semantic information can be added to these datasets by semantic annotation, facilitating more intricate and in-depth data mining and analysis operations. Semantic labeling, for instance, can be used to identify and analyze data about social relationships, user interests, and preferences in social media analysis. In the ecommerce industry, semantic labeling can assist companies in better understanding the demands and purchasing patterns of their customers in order to develop more precise marketing strategies. Data analysis and mining have become essential in the internet era, particularly in scientific research and business decision-making. However, processing large, diverse, and complex internet datasets is a huge challenge for traditional methods. Semantic annotation can inject rich semantic information into the dataset, making data analysis and mining more in-depth and refined. For this reason, research on semantic labeling models and methods based on internet open datasets is of great significance to support data analysis and mining in complex application scenarios.

In this paper, a novel semantic labeling model and method are proposed. The model is based on the internet open dataset and aims to improve the efficiency and accuracy of labeling. The Common Crawl dataset was selected to provide a rich and diverse data basis for the training of the model. In terms of model construction, an architecture based on a two-way short- and long-term memory network is innovatively adopted. This model can fully use the text context, accurately capture the characteristics of parts of speech, and generate corresponding semantic labels for the data through supervised learning. Compared with the traditional semantic labeling model, bidirectional long short-term memory (BiLSTM) shows better feature learning ability and higher labeling accuracy. It can not only cope with large-scale datasets but also ensure the accuracy of labeling while maintaining efficiency. Through this model, we provide a practical solution for processing and applying large-scale internet open datasets. This not only improves the efficiency and accuracy of data labeling but also promotes the sharing and reuse of data, thereby helping innovation and development in various fields.

RELATED WORK

A crucial tool in the field of data processing, semantic annotation models seek to provide semantic information to data (Du, Zhu, et al., 2021; Han et al., 2019). While deep learning models are often seen as more appropriate for semantic tagging, Willrich et al. (2020) felt that supervised learning had largely resolved the problem of semantic tagging. Wang et al. (2021) created a thorough semantic annotation framework for cultural heritage images and its guiding principles and procedures to address the growing demands for semantic enrichment and fine-grained annotation of cultural heritage images. This approach goes beyond Panofsky and information organization theory. According to Liao and Zhao (2019), the gap between a significant amount of unlabeled existing/new data and limited annotation capabilities is the most difficult issue in achieving the semantic web's full potential. Di Martino et al. (2023) developed SemPrAnn, a semantic annotation tool that allowed for the explicit identification of ideas in workflows and the use of reasoning engines to enforce rules. Wahab et al. (2022) suggested a real-world case study and several annotation kinds to investigate algorithm semantics. Fei (2021), based on the need for semantic annotation of resources in the process of resource library construction and sharing, used the latent Dirichlet allocation model to semantically model document resources in the resource library and mine potential topics in the documents.

By including semantic information in the dataset, semantic annotation—a crucial tool in data processing—significantly enhances data comprehension and application effectiveness (Koutsomitropoulos, 2019; Li et al., 2020). Additionally, semantic annotation creates a strong basis for data reuse and sharing, increasing the relevance of systems and applications. Urban areas must