Domain Ontology Evolution Integrating Word Semantic Representation and New Word Discovery: A Case Study of Product Review Data (Postprint)

Authors: Geng Qian, Deng Siyu, Jin Jian

Date: 2023-04-01T00:00:00+00:00

Abstract

[目的/意义] To address the issues of inaccuracy and low efficiency in capturing new knowledge and emerging requirements in traditional ontology evolution, this paper proposes an ontology evolution method based on domain new word discovery, with validation conducted using user product review data as a case study.[方法/过程] First, natural language processing algorithms are employed to preprocess user product review text corpora, and Word2vec is utilized for word vector embedding; then, the Bi-LSTM-Attention-CRF algorithm in deep learning is adopted to identify and extract candidate domain new words, and K-means clustering is applied to obtain the final set of domain new words; finally, the six-stage evolution process of ontology evolution is leveraged to accomplish domain ontology evolution.[结果/结论] Using smartphone domain product reviews as experimental data, the study verifies that the domain new word discovery model achieves higher accuracy and recall rates, thereby evolving a new version of the product ontology for the smartphone domain. This evolved domain product ontology can not only assist product designers in optimizing product design based on new features and functions identified in the ontology, but also support consumers in making purchase decisions by leveraging product reviews.

Full Text

Preamble

Integrating Word Semantic Representation and New Word Identification for Domain Ontology Evolution: A Case Study of Product Review Data

Geng Qian[1],[2], **Deng Siyu**[2], **Jin Jian**[2] [1]Center for Governance Studies, Beijing Normal University, Zhuhai 519087 [2]School of Government, Beijing Normal University, Beijing 100875

Abstract: [Purpose/Significance] To address the problems of inaccuracy and inefficiency in capturing new knowledge and new requirements in traditional ontology evolution, this paper proposes an ontology evolution method based on domain new word identification and validates it using user product review data. [Method/Process] First, natural language processing algorithms were employed to preprocess user product review text corpora, and the Word2vec algorithm was adopted for word vector embedding. Then, a deep learning Bi-LSTM-Attention-CRF algorithm was utilized to recognize and extract candidate domain new words, and the K-means algorithm was applied for clustering to obtain final domain new words. Finally, the six-stage evolution process of ontology evolution was implemented to accomplish domain ontology evolution. [Result/Conclusion] Using smartphone domain product reviews as experimental data, this study verifies that the adopted domain new word identification model achieves higher accuracy and recall rates, thereby evolving a new version of the smartphone domain product ontology. The evolved domain product ontology can help product designers optimize product design based on new features and functions identified in the domain ontology, and also support consumers in making purchase decisions using product reviews.

Keywords: ontology evolution; domain new words; new word detection; attention mechanism; bidirectional long short-term memory network; conditional random field

Classification Number: G250 TP391.1 **DOI:** 10.13266/j.issn.0252-3116.2021.08.009

With the rapid development of e-commerce platforms such as Amazon and Taobao, users can more easily express their opinions on various products. These user reviews can help potential consumers obtain sentiment orientation regarding certain product features to support their purchasing decisions; merchants and product designers can also improve services and enhance product quality based on these reviews [1]. However, user product reviews are unstructured text data with massive content scale [2], containing diverse entities and complex implicit relationships among them [3].

Ontology, as a tool for knowledge organization, has been widely developed and used in numerous practical applications. Utilizing domain ontology enables knowledge organization, storage, and application from user product reviews, providing support for in-depth mining of review content. Knowledge in the real world is constantly updated, and user demands for knowledge are also in a continuous state of change. As products release new features and characteristics, user reviews change accordingly. For example, Apple' s iPhone 11 series released in 2019 featured a new "bath-heater" camera design, while the iPhone X released in 2017 introduced facial recognition functionality. These new features and

functions of mobile products are hot topics that users focus on in their reviews.

The entity words and feature words appearing in user reviews may not exist in the existing domain product ontology, requiring evolution of the original domain ontology to meet new demands. Ontology Evolution refers to the process of modifying existing ontologies to adapt to new knowledge and changing requirements [4]. This evolution reflects the sustainability of ontologies and facilitates their long-term value in practical applications. Moreover, the large scale and continuous development of ontologies make ontology evolution a complex and time-consuming task [5-6].

In ontology evolution, change capturing is the core step in the entire process. If cutting-edge algorithms can be used to automatically identify and extract domain new words from new review text corpora, and these domain new words are used as the changes to be captured in ontology evolution, it would be of great significance for constructing domain product ontology evolution from review texts. It would enable product designers to grasp in real-time the popular functions and components that consumers focus on in product reviews to optimize product design, and also provide support for consumers to make purchase decisions using novel user product reviews. Therefore, this study introduces domain new word identification technology into domain ontology evolution.

Domain new word identification differs from traditional new word identification. Words identified as "new" may have never appeared in a specific domain but not necessarily in all domains. Discovering domain new words can uncover the latest development trends in that domain. For instance, identifying domain new words from user reviews of a certain product category can help people understand the latest functions, components, and packaging currently emerging for that product category. Nowadays, deep learning technology in neural networks has attracted much attention and developed rapidly. Deep learning methods are used to learn the intrinsic rules and representation levels of sample data and discover hidden patterns in data [7]; they can effectively improve the processing effect of word segmentation systems on web text.

Some scholars have also conducted new word identification based on rule-based methods. For example, Zhou Shuangshuang [20] fused rule-based and statistical methods for microblog new word identification; Wang Xin [21] used association rules to rank hot topics in online news; Chen Meijie [22] employed word boundary screening rules when utilizing bidirectional aggregation degree, thereby improving patent new word identification performance. Additionally, some scholars have introduced cutting-edge algorithms and models into new word identification research. Among them, Zhang Huaping et al. [23] used conditional random fields to predict new words in large-scale social media corpora, achieving faster speed and higher accuracy; Wang Ting et al. [24] fused conditional random fields and support vector machines for new word identification, achieving higher precision and recall rates when extracting entities from Chinese encyclopedia classification pages; Chen Xianlai [25] used mutual information and logistic regression algorithms when utilizing new word identification

to improve existing word segmentation models, improving the accuracy of medical text word segmentation; Liu Yutong [26] improved the Apriori algorithm in research on discovering new words from large-scale classical Chinese corpora and added long short-term memory networks and conditional random field algorithms, which was verified to identify new words in Song poetry and Song history datasets; Zhao Zhibin [27] applied syntactic analysis and word vectors to new word identification research, verifying through real text datasets from skincare forums that the method has good performance for new word identification; Huang Wenming [28] used information volume and Bi-LSTM+CRF algorithms for domain new word identification, verifying through Lenovo customer service Q&A system question datasets that the method can improve domain new word identification accuracy.

In summary, although there have been some studies on ontology evolution and new word identification both domestically and internationally, research combining the two is relatively scarce. First, ontology evolution is a series of modification processes made to an ontology while ensuring its consistency, which can be seen as the result of a series of operations in the ontology development process. Research on ontology evolution can be generally categorized into three types: manual evolution methods, semi-automated evolution methods, and automated evolution models or systems.

In manual evolution research, when new knowledge or new requirements emerge, V.S.K. Nagireddi et al. [9] and X. Chen et al. [10] utilized domain experts for evolution or merged existing ontologies with other domain ontologies. In semi-automated evolution research, Liu Ziyu et al. [11] proposed a semi-automated domain ontology evolution method based on DBpedia; Chen Jing et al. [12] optimized the calculation of ripple-effect in ontology evolution based on the SPF algorithm using adjacency lists and evaluated larger-scale ontologies using the Floyd-Warshall algorithm. In automated evolution research, Liu Yi et al. [13] proposed an ontology evolution-driven semantic search engine system—OESSE, which organically combined ontology automatic evolution functionality with semantic search; Liu Ying [14] integrated the social nature of knowledge management into application technology, proposing a distributed knowledge management system based on the linked development of knowledge ontology evolution and information retrieval; C. Huang et al. [15] proposed an ontology generation and evolution system for intelligent manufacturing application goals, which could automatically extract ontologies from raw production data and dynamically adjust ontologies according to changes in the manufacturing data environment.

Second, domain new word identification is the process of recognizing and extracting words that have never appeared before in a specific domain. In traditional new word identification research, mutual information and adjacent entropy have been introduced into new word identification studies [16-18]. Du Liping [19] proposed an improved mutual information algorithm based on the PMIk algorithm and a small number of basic rules, verifying that new word

identification can improve existing Chinese word segmentation systems. Some scholars have conducted new word identification based on rule-based methods, such as Zhou Shuangshuang [20] who fused rule-based and statistical methods for microblog new word identification, Wang Xin [21] who used association rules to rank network news hotspots, and Chen Meijie [22] who employed word boundary screening rules when utilizing bidirectional aggregation degree to improve patent new word identification performance. Additionally, some scholars have introduced cutting-edge algorithms and models into new word identification research. Among them, Zhang Huaping et al. [23] used conditional random fields to predict new words in large-scale corpora in the social media domain, achieving faster speed and higher precision; Wang Ting et al. [24] fused conditional random fields and support vector machines for new word identification, achieving higher precision and recall rates when obtaining entity recognition from Chinese encyclopedia classification pages; Chen Xianlai [25] used mutual information and logistic regression algorithms when utilizing new word identification to improve existing word segmentation models, improving the accuracy of medical text word segmentation; Liu Yutong [26] improved the Apriori algorithm in research on discovering new words from large-scale classical Chinese corpora and added long short-term memory networks and conditional random field algorithms, which was verified to identify new words in Song poetry and Song history datasets; Zhao Zhibin [27] applied syntactic analysis and word vectors to new word identification research, verifying through real text datasets from skincare forums that the method has good performance for new word identification; Huang Wenming [28] used information volume and Bi-LSTM+CRF algorithms for domain new word identification, verifying through Lenovo customer service Q&A system question datasets that the method can improve domain new word identification accuracy.

In summary, although there have been some studies on ontology evolution and new word identification, existing research basically does not combine new word identification methods when evolving large-scale ontology corpora, resulting in poor evolution performance. Second, most domain new word identification studies do not incorporate cutting-edge algorithms and models from deep learning, making it difficult to accurately and quickly discover domain new words when processing large-scale, complex data. Additionally, in research on knowledge organization of unstructured text data, the constructed ontologies often lack longevity and real-time capability, and with minimal later maintenance by developers, they basically fail to sustain their value over time. Therefore, this study will achieve evolution of domain ontologies constructed from user product reviews through ontology evolution technology based on domain new word identification, thereby fully realizing the application value of domain ontologies and providing support for consumers to use product reviews for purchase decisions.

2.2 Current Status of Key Technologies

2.2.1 LSTM Network Long Short-Term Memory (LSTM) network is a type of Recurrent Neural Network (RNN) that has the ability to memorize data sequences. RNN [29] mainly consists of an input layer, hidden layer, and output layer, with the function of memorizing current input and previous input information, and performs better when processing short time series text sequences. However, when processing long time series information, RNN may encounter the problem of gradient vanishing or explosion. Therefore, the LSTM neural network proposed by A. Graves [30] solves the problems existing in RNN and is widely used in image processing, speech recognition, and other fields.

Compared with RNN, LSTM adds memory units and three types of gate structures composed of input gates, forget gates, and output gates on its structural basis [31], as shown in Figure 1 [Figure 1: see original paper]. The gate structure is a fully connected layer in the neural network, where the input vector is processed by the gate structure and outputs a real number vector between 0 and 1. This gate structure of LSTM is based on the sigmoid function, enabling the neural network to have the function of allowing data to pass through (selectively retaining) or discarding state values, facilitating the acquisition of text sequences in long-term sequence distances.

In addition, when predicting text sequences, some prediction results may be jointly determined by several preceding inputs and several following inputs, leading to the emergence of Bidirectional Long Short-Term Memory network (Bi-LSTM), similar to Bidirectional Recurrent Neural Network (BRNN). Bi-LSTM mainly includes two processes: forward propagation and backward propagation. The training sequence is input into the forward LSTM network model, and forward feature information is obtained through forward propagation calculation. Similarly, the backward LSTM network model is input, and backward feature information is obtained through backward propagation calculation. Then the forward feature information and backward feature information are concatenated to obtain the final hidden state, which aggregates bidirectional semantic features from both forward and backward directions. Using Bi-LSTM to solve complex text sequences, such as in user product reviews, since text sequence prediction depends not only on some preceding input texts in the sequence but also on the influence of subsequent input texts, Bi-LSTM can be adopted to improve the prediction accuracy of review texts.

2.2.2 Attention Mechanism The traditional Encoder-Decoder Model is mainly used for processing text, speech, images, videos, and other data, from which algorithms such as RNN and LSTM are derived. When processing text sequences, the encoder encodes the input text sequence into a fixed-length hidden vector and assigns the same weight to the hidden vector. The decoder decodes the output based on these hidden vectors. When the input sequence text content expands and the corresponding component weights of the text sequence are the same, the Encoder-Decoder model' s discriminative power for input

text sequences decreases, causing the model performance to decline accordingly. Therefore, D. Bahdanau et al. [32] proposed the Attention mechanism that can well solve this defect.

The Attention mechanism is used to improve the Encoder-Decoder model effect, quickly screening out high-value information from massive information. Its essence is to simulate human attention, imitating the brain' s thinking activities when humans observe objects [33]. Therefore, the Attention mechanism has important application value in multiple research fields such as sentiment classification and machine translation. In the optimization of the Encoder-Decoder model, the Attention mechanism is mainly used in the decoding process. It changes the traditional Decoder' s disadvantage of assigning the same vector to each input text sequence, but instead assigns different weights according to different words. In the Encoder process, the output is no longer a fixed-length intermediate semantics but a text sequence composed of vectors of different lengths. The Attention mechanism enables the model to assign corresponding weights to the hidden vectors of different moments of the input text sequence, and merges the hidden vectors into new hidden vectors according to their importance, which are finally input to the Decoder. The Decoder process further screens and processes based on this subset of the sequence. Therefore, the Encoder-Decoder model introducing the Attention mechanism is shown in Figure 2 [Figure 2: see original paper].

2.2.3 CRF Sequence Labeling Conditional Random Field (CRF) combines the characteristics of maximum entropy models and hidden Markov models, and is an undirected graph model. The CRF model provides a conditional probability distribution model for another set of output random variables Y given a set of input random variables X, and has been applied to different prediction tasks in sequence labeling [34]. The basic flow of the CRF model is shown in Figure 3 [Figure 3: see original paper].

Before implementing CRF sequence labeling, it is necessary to manually annotate original corpus information and artificially define attributes such as part of speech, degree, and category of words in the corpus. At present, when performing some natural language processing tasks, such as named entity recognition, neural network models are used to learn training data and generate feature vectors to obtain better prediction effects. However, neural network models are relatively time-consuming, and some output results of the model are incorrect recognition results. Therefore, the CRF model can be used for named entity recognition tasks by adding some manually predefined rules to the sequence labeling process, which can achieve better prediction effects.

In summary, LSTM memory units and gate structures effectively solve the gradient vanishing defect in traditional RNN. The bidirectional LSTM model can not only identify past text sequence information but also fully consider future sequence information, making contextual information fully and completely utilized. Introducing the Attention mechanism into the Encoder-Decoder model

well solves the weight of each sequence part when the text sequence length expands. CRF sequence labeling focuses on the linear weighted combination of local features of the entire text sequence, that is, it scans the entire sentence through feature templates and calculates joint probabilities to optimize the entire sequence. Therefore, when processing large-volume user review text, Bi-LSTM neural networks, Attention mechanism, and CRF sequence labeling can be introduced to achieve more accurate text entity recognition effects.

3 Ontology Evolution Based on New Word Identification

This study proposes an ontology evolution framework based on domain new word identification, as shown in Figure 4 [Figure 4: see original paper]. Its core lies in adding domain new word identification to ontology evolution to capture changes in the ontology. Domain new word identification mainly adopts the deep learning model combining Bidirectional Long Short-Term Memory neural network with Attention mechanism and Conditional Random Field (Bi-LSTM+Attention+CRF).

3.1 New Word Identification Based on Deep Learning

As mentioned above, numerous studies have proposed different new word identification methods, such as fusion of information volume [28], mutual information [18-19], syntactic analysis [27], and rules [22, 24]. To further improve the accuracy of new word identification, this study mainly adopts cutting-edge algorithm models in deep learning, such as the Word2vec algorithm, Bi-LSTM-Attention-CRF model, etc. According to the basic flow in Figure 4, new word identification based on deep learning mainly includes five steps: text preprocessing, syntactic analysis, word vector embedding, model training and prediction, and feature clustering. Therefore, this study proposes a new framework for new word identification for ontology evolution, as shown in Figure 5 [Figure 5: see original paper].

3.1.1 Text Preprocessing This study uses user product review data from the Amazon e-commerce platform (Amazon.com) as research data. Before new word identification, it is necessary to preprocess the original English corpus to remove abnormal characters and punctuation from the original text corpus. This study uses Python as the development language and adopts Python's NLTK toolkit for sentence segmentation, word tokenization, eliminating punctuation, removing stop words, POS tagging, etc. Based on POS tagging, words in sentences are normalized through stemming and lemmatization. Finally, original words with part-of-speech annotations are obtained.

This study selected two datasets as the training set and test set. The training set is mainly used to train the Bi-LSTM-Attention-CRF model, while the test set mainly comes from new text corpora for domain new word identification and extraction.

3.1.2 Syntactic Analysis On the basis of preprocessing the original corpus, introducing syntactic analysis can achieve the first screening and extraction of domain candidate new words. Syntactic analysis (Syntactic Parsing), as one of the key low-level technologies in natural language processing, analyzes the grammatical functions of words in sentences [3]. Syntactic analysis is divided into syntactic structure analysis and dependency parsing. To obtain the syntactic structure or complete phrase structure of the entire sentence, it is called syntactic structure analysis; while to obtain local components, it is called dependency parsing.

In user product review texts, domain new words are mainly feature words composed of nouns and verbs, as well as relationship words between feature words. Therefore, to obtain these components in text sentences, this study adopts dependency parsing. Dependency parsing reveals the syntactic structure by analyzing the dependency relationships between components within language units. Intuitively, dependency parsing identifies grammatical components such as "subject-verb-object" and "attribute-adverbial-complement" in sentences and analyzes the relationships between these components. For example, in "Wireless charging damages battery health," "Wireless charging" is the subject, "damages" is the predicate, and "battery health" is the object. Both the subject and object here may become domain new words and can thus serve as candidate domain new words. Therefore, through syntactic analysis, a candidate set of words that may become domain new words in text sentences can be obtained, and manually defined screening rules can be applied to filter out first-stage candidate new words.

3.1.3 Word Embedding In deep learning, using word embedding for feature learning is an effective way to extract entities [28]. Before using deep learning models for data training and prediction, the first task is word embedding, which is the process of converting preprocessed words into numerical vectors, i.e., word vectorization.

Word2vec (Word to Vector) is an open-source deep learning tool for calculating word vectors based on neural network language models and log-bilinear models [35]. By learning text, it captures semantic information of words in the text and represents words in the form of word vectors. Word2vec mainly includes two models: CBOW and Skip-Gram. The CBOW model predicts target word information based on given context, while the Skip-Gram model predicts words that appear in its context based on a given word.

Considering that user product reviews may have multiple complex and potential features and relationships between features—for example, in user review data "It' s not cool for me that AppStore occupy much storage, and waste battery quickly" —there are interwoven relationships among feature words "AppStore," "storage," and "battery," and the association relationships for a feature word can be obtained in the context. Therefore, to improve the accuracy of domain new word identification, this study adopts the CBOW model in Word2vec for word em-

bedding and uses Negative Sampling to reduce training complexity, achieving prediction of word vector representation for a certain vocabulary by training the context of review texts.

3.1.4 Bi-LSTM-Attention-CRF Model Training Based on word embedding, considering that the research data in this study is large-scale unstructured review text with complex and diverse feature entities and potential association relationships, traditional named entity recognition methods have deficiencies in both efficiency and effectiveness. Moreover, deep learning algorithms, such as Bi-LSTM+CRF algorithms, are often more accurate in obtaining domain new words [28]. Therefore, this study introduces the Bi-LSTM-Attention-CRF model, which combines bidirectional long short-term memory neural network with attention mechanism and conditional random field model. The framework of the Bi-LSTM-Attention-CRF model introduced in this study is shown in Figure 6 [Figure 6: see original paper].

The specific implementation of the Bi-LSTM-Attention-CRF model is as follows: by preserving the intermediate output of the Encoder encoder of the bidirectional LSTM for the input text sequence, a model is trained to selectively learn these text inputs and associate the output sequence with them during model output. Then, the bidirectional LSTM plus Attention mechanism learns the features of forward and backward information in the input text sequence, and the CRF model is used to infer the corresponding state sequence based on the given observation sequence, which can utilize the relationship between adjacent front and back labels to obtain the current optimal labeling.

In this study, to discover domain new words in input text, the Bi-LSTM-Attention-CRF model is adopted to process the preprocessed training set and test set. First, the word vector of each word in the input text sequence and the automatically annotated dataset are input. Second, the model is trained on domain word vectors in the training set. Then, the original training set after word segmentation is manually annotated for domain new words using annotation labels of five types: BIE-SO (B_{new}, I_{new}, E_{new}, S_{new}, O). Here, B stands for Begin, indicating the beginning of a new word phrase; I stands for Intermediate, indicating the middle of a new word phrase; E stands for End, indicating the end of a new word phrase; S stands for Single, indicating a single new word character; O stands for Other, used to label irrelevant characters that are not new words. An example of dataset annotation is shown in Table 1 .

Finally, the Bi-LSTM-CRF network is trained using the training set, where each iteration requires forward and backward propagation of Bi-LSTM, encoding and decoding of the Attention layer, and forward and backward propagation of the CRF layer. The trained model is then used to predict domain new words in the test set. After predicting domain new words in the test set data through the Bi-LSTM-Attention-CRF model, second-stage domain candidate new words are obtained.

3.1.5 Feature Clustering Since user product reviews have different expressions for the same feature, function, or component of a product—for example, for the new function of the iPhone X series—facial recognition, some users may use phrases such as "face recognition," "facial recognition," "face scanning," "facial scan," etc. At this point, it is necessary to filter the candidate domain new words identified in the second stage and screen out the real domain new features for ontology evolution. This study adopts high-frequency words among synonyms as the domain new words for that feature.

Additionally, for the domain candidate new words obtained in the second stage, these new words need to be categorized to determine which position in the product ontology they belong to, which facilitates change capture in later ontology evolution work. This study adopts the K-means clustering method in feature extraction. The basic idea of the K-means algorithm is to first determine a constant K, which means the final number of clustering categories. Initially, random points are selected as centroids, and each sample is assigned to the most similar class by calculating the similarity between each sample and the centroid. Then the centroid of each class is recalculated, and this process is repeated until the centroid no longer changes, finally determining the class to which each sample belongs and the centroid of each class. Through K-means, candidate domain new words can be preliminarily clustered, and then domain expert knowledge is used for category judgment to determine domain new features under different categories of a product. Table 2 shows an example of new feature clustering results for iPhone mobile phone products.

3.2 Ontology Evolution Based on New Word Identification

According to the basic flowchart in Figure 4, the basic framework of ontology evolution selected in this study is the six-stage division method proposed by L. Stojanovic et al. [37] in ontology evolution research, which includes change capturing, representation, semantics of change, implementation, propagation, and validation.

The ontology evolution framework proposed in this study builds upon the six-stage division method. First, new word identification technology is added to the change capturing process, as described in Section 3.1. Since the object of this study is review text, and text data changes over time and space, many new features and functions of a certain domain will appear. User product reviews will contain the latest functions and components of new products, such as wireless charging and no Home button that appeared in the iPhone X. Therefore, using new word identification technology can effectively capture changes in concepts, relationships, and instances in the ontology.

Second, representation is the preparatory work for handling changes, essentially using formal methods to represent changes in the domain ontology, including adjustments to the domain ontology structure and concepts, such as using typical domain feature words in product reviews to represent certain aspects of

product features and functions. Semantics of change involves semantic control of domain ontology changes, including changes in concepts, relationships, and instances. In domain product ontologies, concept changes are mainly reflected in adjustments to ontology classes. For example, the Jack class in the original domain ontology contained two subclasses: headphone jack and charging jack, while new products (such as iPhone 8 series, iPhone X series) merged the charging jack and headphone jack into one shared jack, requiring adjustment of this concept to charging and headphone jack.

Relationship changes are mainly reflected in that one-to-one relationships in the original ontology may be adjusted to one-to-many, many-to-one, or even many-to-many relationships. For example, the Price class of new products may be determined by multiple classes, i.e., place of origin, storage, color, screen size, etc. will determine the price of the mobile phone product. Instance changes mainly involve the emergence of some new instances, such as the camera_{pixel} class appearing with single 12-megapixel, rear dual 12-megapixel, etc. Additionally, the body_{color} of iPhone 11 series products has six color instances: purple, white, green, yellow, black, and red.

Implementation includes adjustments to ontology structure and instances. Among them, adjustments to ontology structure include addition, deletion, and modification of classes, object properties, and data properties. For instance adjustments, this study mainly adopts the method proposed by C. Huang et al. [15] for adding and adjusting restrictions on instances. Table 3 shows examples of adjustments to classes and instances in the domain product ontology.

Propagation ensures and maintains the consistency of related ontologies after a domain ontology evolves, to avoid one of the important impacts caused by ontology evolution—incompatibility between previous and later ontology versions. This study uses evolution plugins in protégé, such as the Change-management plugin and PROMPT plugin [38], to propagate and transfer these changes for reuse and inheritance of the new version of this domain ontology by other domain ontologies.

Finally, the validation stage is the final confirmation of the above domain ontology evolution process. After verification through domain experts or machine identification methods for all these steps, the modifications to the domain ontology are confirmed, and some changes can also be deleted according to user needs mined from the text to complete the final confirmation of changes.

4 Experiments and Results Analysis

4.1 Data Source and Preprocessing

This paper conducts experimental research using user product reviews in the smartphone domain as an example. Based on the smartphone domain product ontology constructed in previous research [2-3], this experiment selected

Apple' s iPhone smartphone new product reviews from 2019 on the Amazon e-commerce platform (Amazon.com) as research data to evolve the domain ontology constructed in previous research. A total of 10,437 reviews of the 2013 iPhone 5C/5S series and 2,798 reviews of the 2019 new iPhone 11 series were crawled as experimental data for this study. The 2013 iPhone 5C/5S series mobile phone review data served as the training set, while the 2019 new iPhone 11 series reviews served as the test set data.

Using the text preprocessing method described in Section 3.1.1, the new corpus was processed with sentence segmentation, word tokenization, punctuation removal, stop word removal, POS tagging, and normalization. The Word2vec model was used to generate the word vector space for the new corpus. Before training the Bi-LSTM-Attention-CRF model, manual sequence labeling was performed on the training set and test set using the method shown in Table 1 in Section 3.1.4, with a total of 23,672 words labeled.

4.2 Evaluation Metrics

The domain new word identification technique proposed in this paper can identify new features in the product domain for change capturing in domain ontology evolution. Therefore, when evaluating domain new word identification, the main idea is to compare the smartphone domain new words identified by the Bi-LSTM-Attention-CRF model used in this study with manually annotated domain new words to evaluate the performance of the model used in this study.

The evaluation metrics are: Precision, Recall, and F-measure. The formulas are as follows:

Precision = correct_{{{found}}*{{new}}}{words} / found_{{{new}}_{{words}}}

Recall = correct_{{{found}}*{{new}}}{words} / correct_{{{new}}_{{words}}}

F-measure = 2 × Precision × Recall / (Precision + Recall)

Where correct_{{{found}}*{{new}}}{words} represents the number of domain new words correctly identified by the model; found_{{{new}}*{{words}}} *represents the total number of domain new words identified by the model; correct*{{{new}}_{{words}}} represents the total number of correct domain new words in the new corpus.

4.3 Experimental Results and Discussion

In the domain new word identification experiment, to verify the effectiveness of the method, this study manually annotated 654 groups of domain new words based on Apple' s official documentation, e-commerce platform product details, and domain expert knowledge, serving as correct domain new words in the corpus. The method using the CRF model for filtering new words was used as the baseline, and then compared with the LSTM combined with CRF model,

bidirectional LSTM combined with CRF model, and the Bi-LSTM-Attention-CRF model adopted in this study. The comparison results are shown in Figure 7 [Figure 7: see original paper].

The experimental results in Figure 7 show that the domain new words processed by the Bi-LSTM-Attention-CRF model achieved the best results, with a precision rate of 91.75% and an F-measure of 89.15%.

On the other hand, to verify the universality of the domain new word identification method adopted in this study across different datasets, comparative experiments were conducted on digital camera product review datasets and laptop product review datasets using similar collection and processing methods as the smartphone dataset. The comparison results of identification effects on different datasets under the domain new word identification method using the Bi-LSTM-Attention-CRF model are shown in Figure 8 [Figure 8: see original paper].

The experimental results in Figure 8 show that the accuracy of the domain new word identification method adopted in this study is higher than 85% across different datasets. Due to the relatively fewer new features in digital camera products and laptop products, the recall rate is relatively lower but remains above 70%. Therefore, the above two experimental results can verify that using the model in this study can effectively identify and extract domain new words from review texts.

In the ontology evolution experiment, taking the evolution of the iPhone smartphone domain product ontology as an example, based on the existing domain product ontology from previous research and the results of domain new word identification, the domain ontology can be dynamically adjusted using the ontology evolution method described in Section 3.2. After ontology evolution processing, the old and new versions of the smartphone domain product ontology are presented in Protégé as shown in Figure 9 [Figure 9: see original paper] and Figure 10 [Figure 10: see original paper]. Figure 9 shows the display result of the old version domain product ontology. Figure 10 shows the display result in the OntoGraf module in Protégé, mainly demonstrating the changes in the domain ontology structure after ontology evolution, as well as changes in ontology classes and structure. In Figure 10, the rectangular boxes marked with "new" and bold frames are the new classes added in the new version ontology.

The above results of the new version domain ontology show that previous research was mainly based on early user review text data, and the old version domain ontology constructed accordingly had limited application defects. That is, with changes in time and space, new features and functions emerge in the domain, requiring adjustments to the domain ontology structure. The new version domain ontology evolved under the background of real-time, latest user product review data is reliable when processing new texts. It can help product designers optimize product design based on new features, functions, and components that users focus on in the domain ontology, and also provide support for consumers

when using product reviews to make product purchase decisions, thereby continuing and further exerting the application value of domain product ontologies under product reviews.

References

[1] JIN J, LIU Y, JI P, et al. Review on recent advances in information mining from big consumer opinion data for product design[J]. Journal of computing and information science in engineering, 2019, 19(1): 1-19.

[2] Deng Siyu, Geng Qian, Jin Jian, et al. Construction and application of domain knowledge base based on product review analysis[J]. Information studies: theory & application, 2019, 42(11): 115-122, 127.

[3] GENG Q, DENG S, JIA D, et al. Cross-domain ontology construction and alignment from online customer product reviews[J]. Information sciences, 2020, 531: 47-67.

[4] CARDOSO S D, SILVEIRA M D, PRUSKI C. Construction and exploitation of a historical knowledge graph to deal with the evolution of ontologies[J]. Knowledge-based systems, 2020, 194(22): 105508.

[5] Chen Jing, Liu Zhao, Gu Jinguang, et al. Ripple-effect analysis based on TFOF in ontology evolution[J]. Journal of Wuhan University (natural science edition), 2020, 66(2): 197-204.

[6] BENOMRANE S, SELLAMI Z, AYED M B. An ontologist feedback-driven ontology evolution with an adaptive multi-agent system[J]. Advanced engineering informatics, 2016, 30(3): 337-349.

[7] CHEN C, LIU Y, KUMAR M, et al. Energy consumption modelling using deep learning embedded semi-supervised learning[J]. Computers & industrial engineering, 2019, 135: 757-765.

[8] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[9] NAGIREDDI V S K, MISHRA S. An ontology based cloud service generic search engine[C]//International conference on computer science & education. Colombo: IEEE, 2013: 335-340.

[10] CHEN X, CHEN H, BI X, et al. BioTCM-SE: A semantic search engine for the information retrieval of modern biology and traditional Chinese medicine[J]. Computational and mathematical methods in medicine, 2014, 13(2): 1-13.

[11] Liu Ziyu, Yang Yujia, Zhang Xiaoming, et al. Research on domain ontology evolution method based on DBpedia[J]. Journal of intelligence, 2017, 36(6): 160-166.

[12] Chen Jing, Liu Zhao, Gu Jinguang, et al. Research on optimization of ripple-effect calculation in ontology evolution[J]. Application research of computers,

2020, 37(8): 2366-2370.

[13] Liu Yi, Wang Yu, Yang Deli. Research on personalized semantic search driven by ontology evolution[J]. Journal of the China Society for Scientific and Technical Information, 2015, 34(10): 1048-1055.

[14] Liu Ying. Knowledge management system based on linkage of ontology evolution and knowledge retrieval[J]. Information science, 2016, 34(4): 62-67.

[15] HUANG C, CAI H, XU L, et al. Data-driven ontology generation and evolution toward intelligent service in manufacturing systems[J]. Future generation computer systems, 2019, 101: 197-207.

[16] Liu Weitong, Liu Peiyu, Liu Wenfeng, et al. New word identification based on mutual information and adjacent entropy[J]. Computer engineering and design, 2019, 40(7): 1903-1907, 1914.

[17] Guo Li, Zhang Hengxu, Wang Jiaqi, et al. New word identification algorithm based on Trie tree, word left-right entropy and mutual information[J]. Computer applications, 2019, 36(5): 1293-1296.

[18] Wang Yu, Xu Jianmin. Hot new word identification for network news hotspot recognition[J/OL]. Computer applications: 1-9[2020-09-12]. http://kns.cnki.net/kcms/detail/51.1307.TP.20200722.1337.002.html.

[19] Du Liping, Li Xiaoge, Yu Gen, et al. New word identification based on improved mutual information algorithm for Chinese word segmentation system improvement[J]. Acta scientiarum naturalium Universitatis Pekinensis, 2016, 52(1): 35-40.

[20] Zhou Shuangshuang, Xu Jin' an, Chen Yufeng, et al. Microblog new word identification fusing rules and statistics[J]. Computer applications, 2017, 37(4): 1044-1050.

[21] Wang Xin, Wang Yu, Wang Liang. Network news hotspot ranking based on new word identification[J]. Library and information service, 2015, 59(6): 68-74.

[22] Chen Meijie, Xie Zhenping, Chen Xiaoqi, et al. A new method for extracting bidirectional aggregation degree features for patent new word identification[J]. Computer applications, 2020, 40(3): 631-637.

[23] Zhang Huaping, Shang Jianyun. Open domain new word identification for social media[J]. Journal of Chinese information processing, 2017, 31(3): 55-61.

[24] Wang Ting, Ji Fujun, Xu Tiansheng. A knowledge acquisition method for unstructured information in Chinese network encyclopedia[J]. Library and information service, 2016, 60(13): 126-133.

[25] Chen Xianlai, Han Chaopeng, An Ying, et al. New word identification based on mutual information and logistic regression[J]. Data analysis and knowledge discovery, 2019(8): 105-113.

[26] Liu Yutong, Wu Bin, Xie Tao, et al. New word identification method based on classical Chinese corpus[J]. Journal of Chinese information processing, 2019, 33(1): 46-55.

[27] Zhao Zhibin, Shi Yuxin, Li Binyang. Domain new word identification method based on syntactic analysis and word vector[J]. Computer science, 2019, 46(6): 29-34.

[28] Huang Wenming, Yang Liuqingqing, Ren Chong. Domain new word identification combining information volume and deep learning[J]. Computer engineering and design, 2019, 40(7): 1903-1907, 1914.

[29] GREGOR K, DANIHELKA I, GRAVES A, et al. DRAW: a recurrent neural network for image generation[C]//ICML' 15: proceedings of the 32nd international conference on international conference on machine learning. Lille: JMLR, 2015, 37: 1462-1471.

[30] GRAVES A. Supervised sequence labeling with recurrent neural networks[M]//Studies in computational intelligence, SCI 385. Berlin: Springer, 2012: 5-13.

[31] PALANGI H, DENG L, SHEN Y, et al. Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval[J]. IEEE/ACM transactions on audio, speech, and language processing, 2015, 24(4): 694-707.

[32] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2020-09-16]. https://arxiv.org/pdf/1409.0473.pdf.

[33] Zhang Huali, Kang Xiaodong, Li Bo, et al. Chinese electronic medical record named entity recognition combining attention mechanism with Bi-LSTM-CRF[J]. Computer applications, 2020, 40(S1): 98-102.

[34] Li Gang, Pan Rongqing, Mao Jin, et al. Chinese electronic medical record entity recognition integrating BiLSTM-CRF network and dictionary resources[J]. Journal of modern information, 2020, 40(4): 3-12, 21.

[35] MIKOLOV T. Distributed representations of words and phrases and their compositionality[J]. Advances in neural information processing systems, 2013, 26: 3111-3119.

[36] Hu Tiantian, Dan Yabo, Hu Jie, et al. News named entity recognition and sentiment classification based on attention mechanism and Bi-LSTM combined with CRF[J]. Computer applications, 2020, 40(7): 1879-1883.

[37] STOJANOVIC L, MAEDCHE A, MOTIK B, et al. User-driven ontology evolution management[C]//Proceedings of the 13th international conference on knowledge engineering and knowledge management. Ontologies and the semantic web. Berlin: Springer-Verlag: 2002, 285-300.

[38] NOY N F, CHUGH A, LIU W, et al. A framework for ontology evolution in collaborative environments[C]//International semantic web conference. Berlin: Springer, 2006.

Author Contributions: Geng Qian: Responsible for developing the paper outline and paper revision; Deng Siyu: Responsible for algorithm design and implementation, and paper writing; Jin Jian: Responsible for proposing the paper idea and paper revision.

Note: Figure translations are in progress. See original paper for figures.