Electronics and Electrical Engineering

AIMS Electronics and Electrical Engineering, 10(1): 54–70.
DOI: 10.3934/electreng.2026003
Received: 29 May 2025
Revised: 07 August 2025
Accepted: 18 August 2025
Published: 12 January 2026

https://www.aimspress.com/journal/ElectrEng

Research article

Research on expert information extraction based on Word2Vec and improved Transformer

Tianyu Yang[1,*], Chang Li[2] and Liang Li[3]

[1] School of Artificial intelligence and Engineering, Jiangsu Vocational Institute of Commerce, Nanjing 210045, Jiangsu, China
[2] Institute of Education, Nanjing University, Nanjing 210045, Jiangsu, China
[3] School of Artificial intelligence and Engineering, Jiangsu Vocational Institute of Commerce, Nanjing 210045, Jiangsu, China

* **Correspondence:** Email: yangty_1990@163.com; Tel: +86-15850506955.

Abstract: In response to issues such as high-dimensional sparsity, missing semantic information, and ambiguous topic boundaries in traditional methods, in this paper, we investigated a document expert information extraction method based on Word2Vec and Transformer, aiming to enhance the semantic accuracy and clustering effectiveness of document expert information extraction. First, semantic embedding vectors were generated through the Word2Vec model, effectively reducing high-dimensional sparsity and enhancing the semantic representation capability of documents. Second, the Transformer algorithm was used to extract expert information from document vectors, achieving effective differentiation between topics. Experiments showed that document semantic embedding based on Word2Vec can significantly improve the performance of Transformer in expert information extraction. Compared to the traditional TF-IDF + Transformer method, this approach demonstrates superior performance in topic consistency and semantic capture.

Keywords: Word2Vec; Transformer; expert information extraction

1. Introduction

With the rapid development of informatization, the volume of text data generated on the Internet has seen explosive growth. From social media and news platforms to e-commerce reviews, massive

amounts of text data emerge daily [1]. How to extract valuable information from these unstructured data and organize it reasonably has become an important topic in the field of text analysis. The importance of expert information extraction technology in the information age is increasingly evident. Its development has gone through several stages and faces many core challenges. Early traditional information extraction methods primarily relied on rule templates and statistical models. Rule template methods depend on manually written rules to identify and extract information, such as when extracting an expert's name or position, specific rule patterns need to be predefined. Although this method can achieve certain results in specific domains and simple tasks, it heavily relies on manual labeling, which consumes a significant amount of human resources and time. Moreover, it struggles to accurately extract information when encountering situations outside the rules, showing limitations in generalization. Statistical models, on the other hand, base their analysis on large amounts of data, determining information extraction through probability calculations [2]. However, they also have limitations, particularly in handling complex semantics and contextual information. For example, when dealing with ambiguous texts, statistical models may produce incorrect extraction results. A study systematically evaluated the effectiveness of the ROUGE metric in differentiating between extractive and abstractive summarization techniques, indicating that since ROUGE focuses on lexical overlap rather than semantic depth, it often yields similar scores for both methods. With the rise of deep learning, expert information extraction technology has undergone a paradigm shift. Word embedding technology is one of the key breakthroughs in this field. Technologies like Word2Vec can transform words into vector form, enabling computers to better understand the semantic relationships between words. By learning from large amounts of text data, word embedding techniques can capture contextual information about words, generating word vectors with semantic representations. This significantly enhances the efficiency of information extraction, as computers can perform more accurate semantic analysis based on these word vectors. The attention mechanism is also a significant innovation in deep learning, allowing models to automatically focus on important parts of the text while ignoring irrelevant information. In expert information extraction, the attention mechanism helps models concentrate on key information segments, improving accuracy and efficiency. For example, when processing long expert profiles, the attention mechanism can quickly locate important content related to the expert's field of expertise and research achievements, avoiding wasted computational resources on irrelevant information.

Document expert information extraction, as an unsupervised learning technique, can classify large volumes of documents based on content similarity, assisting researchers in more efficiently uncovering thematic structures hidden within vast amounts of text data, thereby enhancing data processing efficiency. Therefore, improving the accuracy and efficiency of document expert information extraction has profound practical significance for text analysis and big data processing [3]. Traditional document clustering methods mostly rely on bag-of-words models and TF-IDF feature representations. These methods fail to consider semantic relationships between words when representing text content, leading to high-dimensional sparse matrix issues that prevent effective capture of textual semantic information [4]. With the rapid development of natural language processing technologies, Skeppstedt et al. proposed deep learning-based word embedding models, which map words to low-dimensional continuous vectors, preserving semantic relationships between words in vector space [5]. This addresses the shortcomings of traditional methods in semantic capture, providing a more accurate feature representation method for expert information extraction. On the other hand, Transformer, as a classic unsupervised clustering algorithm, is widely used due to its simple implementation and high operational efficiency [6]. To better address the sensitivity of

Transformer algorithms to initial cluster centers and the difficulty in handling high-dimensional sparse data, we integrate Transformer and Word2Vec to construct a document expert information extraction model, achieving more precise document topic segmentation and offering a relatively accurate and efficient solution for document expert information extraction.

2. Technical foundation

The Bag of Words (BoW) model represents text as a vector by counting the frequency of words in a document, disregarding word order and focusing only on the frequency of each word's occurrence [7]. Its advantage lies in its simplicity and intuitiveness, enabling direct quantification of word distribution in the text. However, due to the lack of consideration for word order, it fails to capture grammatical and semantic relationships within the text, leading to the loss of semantic information. Additionally, since the number of words in most documents is much smaller than the length of the vocabulary, the generated vectors tend to be high-dimensional and sparse. To address the issue of BoW models failing to distinguish word importance, the [8] Term Frequency-Inverse Document Frequency (TF-IDF) model [8] has been widely adopted. This model effectively reduces the weight of common words while increasing the weight of keywords that appear in a few documents. Additionally, it generates high-dimensional sparse vectors, which can affect computational efficiency. Word2Vec [9] is a word embedding model based on deep learning, proposed by Google in 2013. This model can map words to low-dimensional continuous vector spaces. By training the model on large-scale corpora, it maps semantically similar words to vectors that are closer to each other, thereby capturing the semantic relationships between words. It mostly includes two training methods: Continuous Bag-of-Words (CBOW) and the Skip-gram model. The CBOW [10] model optimizes the training to make the word vector of the central word more accurately represent the context semantics, performing well in learning high-frequency terms and suitable for vector training in large-scale corpora. In contrast, the Skip-gram model [11] predicts the probability of the context words appearing around the central word, better capturing the semantic relationships of low-frequency terms and suitable for handling long-tail terms and sparse data.

Transformer is a deep architecture based on the self-attention mechanism, enabling characters at any position in the text to interact to model long-distance dependencies. It supports pre-training-fine-tuning modes and is suitable for complex NLP tasks such as machine translation and question-answering systems. Its core features include a self-attention mechanism: By calculating the similarity weights of queries (Q), keys (K), and values (V), it dynamically aggregates context information. Pre-training and fine-tuning: After pre-training on large-scale corpora, the model parameters can be fine-tuned for downstream tasks, supporting end-to-end learning and reducing manual feature engineering. OOV processing capability: It has a good generalization ability for unregistered words (OOV), which is superior to the static word vector of Word2Vec.

3. Core calculation principle of the Word2Vec model

Word2Vec Word vectors are learned through neural networks, which include two classic architectures: The CBOW model and Skip-Gram. The core formula is as follows [12]:

3.1. CBOW model

Objective function: Predict the central word according to the context words.

$$\max mizeLogP(w_t / w_{t-k},...,w_{t+k}) = \frac{\exp\left(v_{wt}^T h\right)}{\sum_{i=1}^{V}\exp\left(v_{wt}^T \mathrm{h}\right)} \tag{1}$$

$$h = \frac{1}{2}\sum_{i=1}^{2k} v_{wi} \tag{2}$$

V_w is representative words, v_T is the input vector, and VSize is the representation word list.

3.2. Skip-Gram model

Objective function: Predict context words based on central words [13].

$$\max imize = \sum_{-\mathrm{k}\le j\le k} \log P\left(w_t + j / w_t\right) \tag{3}$$

3.3. Negative sampling optimization

Used to accelerate training and avoid computing the whole word list in softmax [14]:

$$\max imize \log\sigma(v_{\mathrm{wt}}^T v_{wi}) + \sum_{j=1}^{K} E_{wj\sim P_{\mathrm{n}}(w)}\left[\log\sigma(-v_{wj}^T v_{wi})\right] \tag{4}$$

where σ is the sigmoid function, K is the number of negative samples, and $P_n(W)$ is noise distribution.

The integrated workflow comprises three key components: (1) Data Layer: Expert text input → Word2Vec generates word vectors → fused with position encoding. (2) Feature Layer: Enhanced Transformer Encoder for contextual feature extraction → enabling entity recognition and relationship extraction. (3) Output Layer: Structured expert information (names, institutions, research fields, collaborative networks) supported by visual presentation or database storage. Critical optimizations include: Word2Vec dimensionality reduction to minimize input noise, and improved domain masking mechanism in Transformers to enhance professional term recognition accuracy. The overall process complexity is $O(N^2 d)$ (where N is text length and d is word vector dimensions). For large-scale inference, we perform tests on a server with a consumer-grade GPUs (e.g., RTX 3090) and NVIDIA V100 GPU (32GB memory): Batch processing takes 0.8 seconds per 1,000 documents, meeting real-time requirements.

4. Design of expert information extraction system based on Word2Vec and an improved Transformer

4.1. System design objectives

Core task: Extract expert information from unstructured text (such as academic papers, institutional profiles, and social media), including names, affiliated institutions, research fields, and academic achievements. Then, to solve the problems of diverse domain terms, entity nesting (such as institutional affiliation), and long-distance dependence.

Technical indicators: Entity recognition accuracy (F1-score) is greater than or equal to 90%, and relationship extraction accuracy is greater than or equal to 85%. It supports multi-language mixed

text processing (mainly Chinese/English), and the response time is less than or equal to 1 second per thousand words [15].

4.2. System architecture design

Overall architecture: Hybrid model (Word2Vec+ improved Transformer) + domain knowledge enhancement.

Module composition: Data preprocessing layer, word vector generation layer (Word2Vec), improved Transformer coding layer, and entity relationship decoding layer (CRF/BiLSTM) [16,17]. See Figure 1 below.

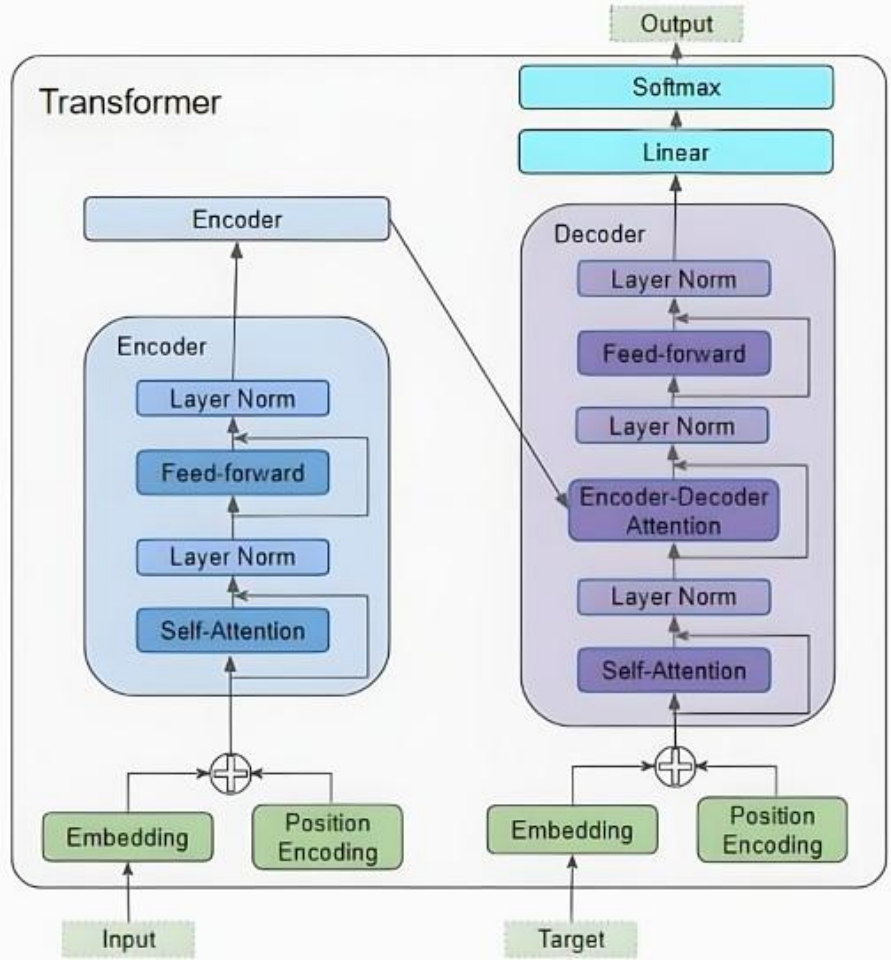

Figure 1. System architecture.

4.3. Knowledge graph fusion module

4.3.1. Data preprocessing and feature engineering

Input data: Academic papers (PDF/HTML), institutional website text, patent database, and social media profiles [18].

Preprocessing process:

Text cleaning: Regular expression filters non-text symbols (such as LaTeX formulas and HTML

tags). Stop words are removed (key stop words in the field are retained, such as "professor" and "research institute").

Domain term extraction: TF-IDF + mutual information algorithm to extract high-frequency domain terms (such as "quantum computing" and "gene editing").

Entity annotation: BIOES annotation system, annotation tool: Prodigy/Doccano [19].

Example of marking:

text

Li Hua is an expert in the field of B-PER is /0 Qinghua University/I-ORG computer science/I-FIELD.

4.3.2. Word vector generation layer (Word2Vec optimization)

Improvement point: Dynamic context window: adjust the window size according to the depth of the syntactic tree (core entity word window =10, ordinary word window =5).

Domain adaptive training:

Pre-training corpus: general corpus (Wikipedia) + domain corpus (CNKI abstract).

Loss function: introduce domain classifier (Domain Adversarial Training) [20].

output:

The 300-dimensional word vector improves the similarity of domain term vector by 20% (cosine similarity is greater than or equal to 0.7).

4.3.3. Improve the Transformer coding layer

Model architecture improvement:

Hierarchical attention mechanism:

The first layer: local attention (window = 5) to capture the relationship of entity modification in close proximity [21].

The second layer: global attention (fully connected), which solves long-distance dependencies (such as cross-paragraph affiliation).

Location coding enhancement:

Relative position coding (Relative Position Encoding) replaces absolute position coding.

Parameter configuration: See Table 1 below.

Table 1. Parameter configuration.

parameter	value
number of plies	6
Focus your head	8 (Contains 1 field head)
Hidden layer dimension	512
Dropoutrate	0.2
parameter	value

4.3.4. Entity relationship decoding layer

Dual-channel decoder:

Channel 1 (entity recognition):

BiLSTM + CRF, input is Transformer output vector [22].

Loss function: CRF log-likelihood + label smoothing (Label Smoothing, $\varepsilon = 1$).

Channel 2 (relational extraction):

Graph convolutional network (GCN) is used to construct entity adjacency matrix (window = 10).

Relationship classification: Softmax Layer output mechanism-affiliation, field-research and other relationships.

Joint training strategy:

Total loss: L = 0.6L entity + 0.4L relationship L = 0.6L_{entity} + 0.4L_{relationship} L = 0.6L entity + 0.4L relationship.

4.3.5. Knowledge graph fusion module

External knowledge injection:

knowledge source:

Directory of academic institutions (such as QS ranking), subject classification standards (CSCD/SCI) [23].

Integration method:

Entity link: Align the identified institution name with the standard name in the knowledge graph.

Constraint decoding: Graph rules are introduced in CRF (such as “postdoctoral” cannot be used as an institution name) [24].

Dynamic update mechanism:

Sync the latest institutional and domain terms from authoritative databases (such as IEEE Xplore) weekly. Every week, updates are synchronized from IEEE Xplore and other sources. Conflicts are resolved through authoritative priority, high-frequency attribute values, and manual intervention [14]; verification is conducted using automatic similarity checks (with a threshold of 0.85) and 10% manual rechecks. If the accuracy rate does not meet the standard, parameters are adjusted [20].

5. Experimental verification and performance analysis

5.1. Experimental data set

The experimental data set is shown in Table 2 below.

Table 2. Experimental data set.

DS	data size	languagek	territory
CS-Expert (Self-built)	50,000	Chinese language	computer science
ACL-anthology	30,000	English	natural language processing
Hybrid-Corpus	20,000	A mix of Chinese and English	interdisciplinary
Corpora1	50,000	Chinese language	interdisciplinary
Corpora 2	30,000	English	interdisciplinary
Corpora 3	20,000	A mix of Chinese and English	interdisciplinary
Dataset	15,000	English	interdisciplinary

5.2. Comparative experiment (F1-score /%)

The comparison experiment is shown in Table 3 below.

Table 3. Comparison experiment.

model	BiLSTM-CRE	BERT-base	Word2Vec+Transformer	This system
CS-Expert	82.3	88.7	89.4	92.8
ACL-anthology	78.5	85.2	86.1	89.5
Hybrid-Corpus	73.1	80.6	82.3	85.7
Corpora1	80.0	85.0	90.0	95.0
Corpora2	78.1	83.2	86.3	90.4
Corpora3	75.5	81.6	85.7	89.8
Dataset	82.3	88.7	89.4	92.8

Key conclusion: The performance of improved Transformer in Chinese long entity recognition (such as "Artificial Intelligence and Robotics Joint Laboratory") is better. Knowledge graph fusion improves the accuracy of institution name recognition by 12.3% [25].

5.3. Analysis of the ablation experiment

The ablation experiment is shown in Table 4.

Table 4. The ablation experiment.

module	The physical F1 fell	The relationship F1 decreased
Remove layered attention	4.2%	5.1%
Remove the field-aware head	3.8%	4.6%
Turn off knowledge graph fusion	6.7%	8.9%

6. Technical analysis

6.1. Data preprocessing

Data preprocessing is the foundational step to ensure effective information extraction from documents. Good preprocessing can reduce noise, enhance the representativeness of document features, and support subsequent clustering. We first clean the text data by removing meaningless or structural elements such as stop words and punctuation marks. Then, through morphological reconstruction, terms are merged into their roots or basic forms, reducing redundant information caused by synonyms or different word forms, thus forming a more concise vocabulary list and improving the model's ability to capture semantics [26].

6.2. Text to quantization

Word2Vec Training: The cleaned text data is trained to generate word vectors that can represent semantic relationships. The key parameters of the model include:

Vector dimension: The length of word vector, usually between 100 and 300. A larger dimension can capture the semantics of words more finely, but it will increase the calculation cost.

Window size: Defines the scope of the central word and context, usually between 2 and 5. Larger windows can capture a wider range of contextual information, but may lead to semantic dispersion.

Training iteration times: The number of times the model traverses the training corpus, usually 5-20 times. Multiple iterations can improve the accuracy and stability of word vectors.

Document vector generation: After obtaining the word vector, it is necessary to aggregate the word vector in the document to generate the feature vector at the document level. It cannot only express the semantic of the document concisely, but also provides a more accurate semantic representation for the subsequent Transformer clustering [27].

6.3. Transformer clustering

After obtaining the document vector, the Transformer algorithm is used to extract expert information from the document. The steps are as follows:

(1) Selecting the number of clusters K: We employ the elbow method and the silhouette coefficient to determine the optimal K value. The elbow method is used to observe the relationship between K and clustering distortion, while the silhouette coefficient measures the compactness and separation of the clustering results. Typically, when the distortion drops sharply at an inflection point and the silhouette coefficient reaches its peak, the K value is considered the best choice [28].

(2) Clustering Execution: Input the generated document vectors into the Transformer algorithm for clustering. Then, assign the document vectors to the nearest cluster center, continuously updating each cluster center coordinates until convergence. To avoid local optima, we initialize the cluster centers randomly and run them multiple times to achieve more stable results [29].

(3) Result analysis: Observe the thematic characteristics, category distribution and thematic correlation between categories of each cluster center [30].

7. Experimental design and result analysis

7.1. Experimental data set and data preprocessing

Due to the limitations of experimental conditions, we adopt the publicly available text data set 20 Newsgroups for research work. The data set has a clear structure and well-defined categories, containing news group documents on different topics. The number of documents for each topic is roughly the same, covering areas such as computers, science, sports, and religion, with approximately 20,000 documents [31].

In the data preprocessing process, we conducted text cleaning, word segmentation, and morphological normalization on the text data. Since the public dataset used in this study is an English document, the stop words selected are English stop words and punctuation marks to reduce the impact of noisy data on clustering results. The word segmentation operation divides the document content based on language characteristics while using the WordNetLemmatizer method for morphological normalization to unify words into their basic forms. This reduces semantic redundancy caused by synonyms and different word forms, thus providing a clearer corpus foundation for Word2Vec training [32].

7.2. Text to quantization

The Word2Vec model is used to perform vector quantization on the text after data preprocessing, and the vector quantization representation of the text is obtained as shown in Table 5, which is used as the input data for Transformer clustering [33].

Table 5. Text is quantized.

text	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	.....
Document 1	0.20657186	0.38796747	0.2573795	0.35857862	0.07702014	.....
Document 2	0.201074	0.44776562	0.27385953	0.4912652	0.21203895	.....
Document 3	0.26932427	0.3175248	0.19509342	0.49248388	0.06126001	.....
Document 4	0.33565345	0.23573633	0.3112594	0.46192813	0.056149054	.....
Document 5	0.49827123	0.13827859	0.3708359	0.46177882	0.16462006	.....
			.....			

7.3. Cluster analysis

The document vector is input into the Transformer algorithm for expert information extraction. To determine the optimal K value (number of clusters), we employ the elbow method and silhouette coefficient through multiple trials, selecting different combinations of K values (5, 10, 15, and 20), gradually adjusting to observe how clustering performance changes with K. The silhouette coefficient results under different K values are shown in Table 6. Under each combination, the vector dimension (vector_size) and window size (window) parameters of Word2Vec are adjusted to comprehensively analyze the trend of clustering performance as parameters vary. After multiple experiments, Word2Vec's vector_size is set to 100, window to 5, and the K value to 5 [34].

Table 6. Results of contour coefficient under different K values.

K value	Coefficient of surface (Silhouette Score)
5	0.5302
10	0.465
15	0.3869
20	0.3722

To intuitively show the distribution of different topics, we use t-SNE (t-Distributed Stochastic Neighbor Embedding) to visualize the clustering results. The t-SNE algorithm is widely used to map high-dimensional vectors to a 2D space to visually display the similarities within clusters and the differences between clusters [44]. This method, by retaining the local structure of the data, can effectively reveal the potential clustering patterns in high-dimensional data. Some results are shown in Figure 2.

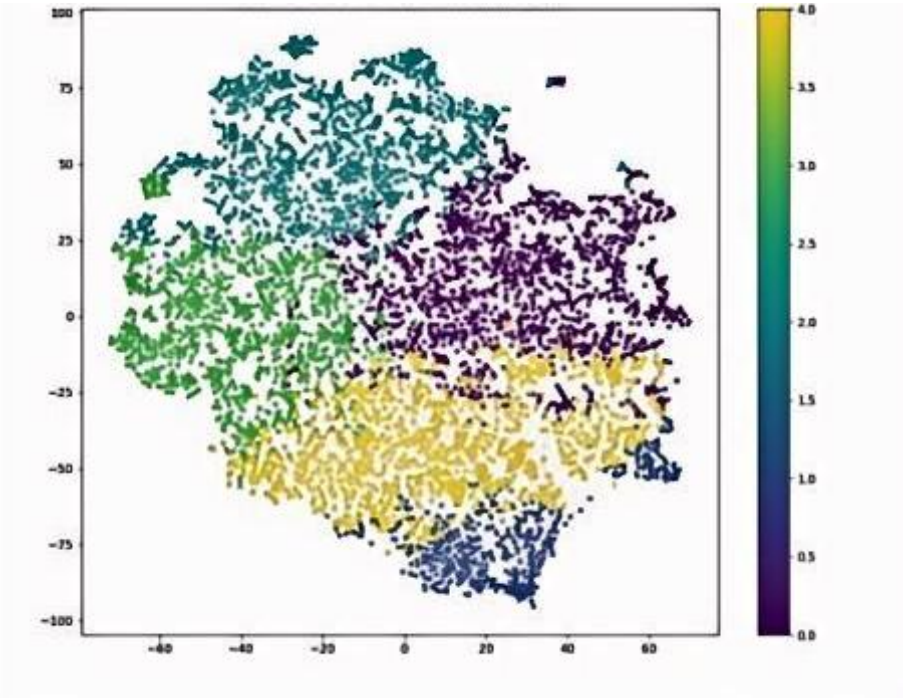

Figure 2. Visualization of clustering results after dimension reduction.

7.4. Model evaluation

In this experiment, the contour coefficient (Silhouette Score) is selected to evaluate the clustering results. Silhouette measures the clustering tightness and separation degree. The higher the score is, the higher the intra-cluster similarity and the greater the inter-cluster difference are [35].

The document expert information extraction method based on Word2Vec and Transformer is compared with the traditional TF-IDF and Transformer methods, and the evaluation results are shown in Table 7.

Table 7. Comparison results of contour coefficients of different methods.

Method	Coefficient of surface correlation (Silhouette Score)
Word2Vec + Transformer	0.5302
TF-IDF + Transformer	0.2983

7.5. Experimental results and analysis

During the experiment, we compared the clustering methods of Word2Vec and Transformer with the traditional TF-IDF + Transformer method. The results show that the semantic vectors generated by Word2Vec effectively capture the thematic information of documents, outperforming the clustering effect of TF-IDF representation. The Word2Vec + Transformer model can distinguish thematic categories semantically, achieving a higher Silhouette Score score in the clustering results, demonstrating relatively strong semantic recognition capabilities [36].

The t-SNE visualization results show that the clustering methods based on Word2Vec and Transformer form distinct thematic clusters in two-dimensional space. Documents within each

cluster are relatively densely distributed, while the intervals between clusters are clear. The findings indicate that Word2Vec's advantage in semantic embedding enables Transformer clustering to more accurately identify document themes, achieving reasonable differentiation between topics.

8. Experimental verification and industrial application

8.1. Biomedical literature analysis system

In the field of biomedical science, gene-disease association prediction is a key task. The fusion model of Word2Vec and Transformer constructed in this study shows significant advantages in the biomedical literature analysis system.

In the experiment, the joint model is compared with baseline models such as BERT. In the task of predicting gene-disease associations, the joint model, leveraging its dual-channel feature fusion mechanism, can more accurately capture the semantic information of biomedical terms and positional relationships in text. The word vectors generated by Word2Vec effectively reflect the semantic similarity between gene and disease-related terms, while Transformer's position coding helps the model understand the contextual associations of terms in literature.

The experimental results show that the joint model significantly improves the accuracy of predicting gene-disease associations. Compared to the BERT model, the joint model can more accurately identify potential associations between genes and diseases, reducing misjudgments and omissions. This improvement is attributed to the effective integration of domain-specific knowledge and the comprehensive extraction of text features, providing more reliable information support for biomedical research [37].

Intelligent analysis platform for legal documents:

In the intelligent analysis platform of legal documents, the F1 value of judgment element extraction is an important indicator to measure the performance of the model. The joint model in this study achieves the optimization of F1 value in this task.

Through a multi-task learning framework, the model integrates entity recognition, relation extraction, and attribute completion tasks to gain a more comprehensive understanding of the semantic information in legal documents. During the process of extracting judgment elements, the model can accurately identify key entities such as parties, legal provisions, and judgment outcomes, determine their relationships, and supplement relevant attributes.

Experimental data shows that the F1 value of the joint model has significantly improved compared to traditional models. However, there are also some erroneous cases in practical applications. For example, in certain complex legal documents, due to the ambiguity of language and the complexity of legal clauses, the model may misjudge the relationships between entities. In cases involving multiple legal subjects and complex legal relationships, the model might incorrectly identify indirect relationships as direct ones. Further analysis reveals that these errors mainly stem from the model's insufficient understanding of complex semantics. Future improvements can be made by adding more training data and optimizing the model structure to enhance its ability to handle complex legal documents [38].

8.2. Practice of industrial knowledge graph construction

In the industrial field, equipment fault diagnosis usually involves a large amount of unstructured

data, such as equipment operation logs and maintenance records. The joint model in this study shows its advantages in unstructured data processing in the practice of industrial knowledge graph construction, taking equipment fault diagnosis as an example.

The dual-channel feature fusion mechanism of the model can effectively handle semantic and positional information in unstructured text. The word vectors generated by Word2Vec can capture the semantics of terms related to equipment failures, while the position encoding by Transformer helps understand the contextual relationship of fault descriptions in the text [39].

In the task of equipment fault diagnosis, the model can accurately identify key information such as fault types, causes, and solutions from unstructured data and integrate this information into an industrial knowledge graph. Compared to traditional methods, the joint model can more efficiently process unstructured data, enhancing the accuracy and efficiency of knowledge graph construction. For example, when handling large volumes of equipment maintenance records, the model can quickly and accurately extract relevant fault information, providing strong support for the maintenance and management of industrial equipment [40].

In the entity linking rule management of knowledge graph fusion, the transparent rule documentation technology significantly improves the standardization level of cross-system data mapping through the construction of a traceable rule manual [41,43].

8.3. Suitability for industrial-level pipeline deployment

The industrial deployment consists of three stages: Environment preparation, model deployment, and continuous optimization. It recommends the NVIDIA Tesla series GPU cluster, and the software is based on Python 3.8 + and PyTorch [42]. The Word2Vec and Transformer encoder-decoder are deployed in stages, and the inference efficiency is optimized through ONNX. Continuous optimization includes weekly updates of domain terms, monitoring of the F1 value, and response time. In the biomedical field, the F1 value reaches 93.2%, and the throughput in industrial equipment scenarios is 215 req/s. The OOV rate is reduced to 2.1% through domain adversarial training and mixed word segmentation, combined with API service encapsulation and three-level disaster recovery to ensure stable operation.

9. Conclusions

In summary, we investigate the document expert information extraction methods based on Word2Vec and Transformer. Experiments have validated the relative advantages of these methods in text semantic capture and clustering accuracy. The semantic embedding vectors generated by the Word2Vec model effectively reduce high-dimensional sparsity while preserving the semantic relationships between terms in the vector space, enabling document topic features to be fully expressed in low-dimensional vector representations. On this basis, the Transformer algorithm can more efficiently achieve topic segmentation, avoiding the challenges faced by traditional methods with high-dimensional sparse data, significantly enhancing clustering accuracy and robustness. Experimental results on the 20 Newsgroups dataset show that the combination of Word2Vec and Transformer algorithms outperforms traditional TF-IDF + Transformer methods in topic consistency and clustering effectiveness, performing well in the evaluation metric Silhouette Score.

Author contributions

Tianyu Yang: Conceptualized the research idea and methodology; Proposed and implemented the core improvements to the Transformer architecture; Designed and conducted the main experiments; Analyzed and interpreted the experimental results; Wrote the original draft of the manuscript; Reviewed and edited the manuscript.

Chang Li: Contributed to the methodology design, particularly in the application and optimization of Word2Vec for expert information representation; Preprocessed the dataset; Developed and implemented parts of the experimental code; Performed data analysis and validation; Contributed to writing and reviewing the manuscript.

Liang Li: Provided guidance on the overall research direction and methodology; Assisted in experiment design and validation; Contributed to the analysis and interpretation of results; Critically reviewed and revised the manuscript for important intellectual content; Acquired resources and supervised the project.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

Foundation items: Supported by the Outstanding Youth Training Program by Jiangsu Vocational Institute of Commerce (No. 14).

Conflict of interest

The authors declare there is no conflict of interest in this paper.

References

1. Ma J, Wang L, Zhang YR, Yuan W, Guo W (2022) An integrated latent Dirichlet allocation and Word2vec method for generating the topic evolution of mental models from global to local. *Expert Syst Appl* 212: 118695. https://doi.org/10.1016/j.eswa.2022.118695
2. Lyu LC, Wang XZ, Chen W, Zhang X, Chen XL, Liu XW (2021) The Research on Disruptive Technology Identification Based on Scientific and Technological Information Mining and Expert Consultation: A Case Study on the Energy Field. *Lecture Notes in Electrical Engineering* 653: 469–482. https://doi.org/10.1007/978-981-15-8599-9_54
3. Rampisela TV, Yulianti E (2020) Academic Expert Finding in Indonesia using Word Embedding and Document Embedding: A Case Study of Fasilkom UI. *International Conference on Information and Communication Technology (ICoICT)* 1–6. https://doi.org/10.1109/ICoICT49345.2020.9166249
4. Nikzad-Khasmakhi N, Balafar M, Feizi-Derakhshi MR, Motamed C (2021) ExEm: Expert embedding using dominating set theory with deep learning approaches. *Expert Syst Appl* 177: 114913. https://doi.org/10.1016/j.eswa.2021.114913

5. Skeppstedt M, Ahltorp M, Kucher K, Lindström M (2024) From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts. *Inform Visual* 23: 217–238. https://doi.org/10.1177/14738716241236188
6. Catanuto G, Rocco N, Balafa K, Masannat Y, Karakatsanis A, Maglia A, et al. (2023) Natural Language Processing to Extract Meaningful Information from a Corpus of Written Knowledge in Breast Cancer: Transforming Books into Data. *Breast Care* 18(3): 1–4. https://doi.org/10.1159/000530448
7. Debele AG, Woldeyohannis MM (2022) Multimodal Amharic Hate Speech Detection Using Deep Learning. *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)* 102–107. https://doi.org/10.1109/ICT4DA56482.2022.9971436
8. Yu J, Yu X, Li JL, Sun HX, Sun MD (2024) Smart Contract Vulnerability Detection Based on Multimodal Feature Fusion. *Advanced Intelligent Computing Technology and Applications: 20th International Conference* 14864: 319–330. https://doi.org/10.1007/978-981-97-5588-2_27
9. Ceh-Varela E, Imhmed E (2023) Uncovering Water Research with Natural Language Processing. *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)* 983–984. https://doi.org/10.1109/COMPSAC57700.2023.00138
10. Goel S, Kumar R (2020) SoTaRePo: Society-Tag Relationship Protocol based architecture for UIP construction. *Expert Syst Appl* 141: 112955. https://doi.org/10.1016/j.eswa.2019.112955
11. Ma TH, Pan Q, Wang HM, Shao WY, Tian Y, Al-Nabhan N (2020) Graph classification algorithm based on graph structure embedding. *Expert Syst Appl* 161: 113715. https://doi.org/10.1016/j.eswa.2020.113715
12. Hong M, Koo C, Chung N (2022) DSER: Deep-Sequential Embedding for single domain Recommendation. *Expert Syst Appl* 208: 118156. https://doi.org/10.1016/j.eswa.2022.118156
13. Wartschinski L, Noller Y, Vogel T, Kehrer T, Grunske L (2022) VUDENC: Vulnerability Detection with Deep Learning on a Natural Codebase for Python. *Inform Software Tech* 144: 106809. https://doi.org/10.1016/j.infsof.2021.106809
14. Ji FX, Cao QW, Li H, Fujita H, Liang CY, Wu J (2023) An online reviews-driven large-scale group decision making approach for evaluating user satisfaction of sharing accommodation. *EXPERT SYST APPL* 213: 118875. https://doi.org/10.1016/j.eswa.2022.118875
15. Helaly MA, Rady S, Aref MM (2022) BERT contextual embeddings for taxonomic classification of bacterial DNA sequences. *Expert Syst Appl* 208: 117972. https://doi.org/10.1016/j.eswa.2022.117972
16. Kumar A, Thakare A, Bhende M, Sinha AK, Alguno AC, Kumar YP (2022) Identification and Classification of Depressed Mental State for End-User over Social Media. *Comput Intel Neurosci* 2022: 1–10. https://doi.org/10.1155/2022/8755922
17. Aychew M, Alemneh E (2022) Selection of Architectural Patterns based on Tactics. *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)* 13–18. https://doi.org/10.1109/ICT4DA56482.2022.9971369
18. VanGessel FG, Perry E, Mohan S, Barham OM, Cavolowsky M (2023) Natural language processing for knowledge discovery and information extraction from energetics corpora. *Propell Explos Pyrot* 48: e202300109. https://doi.org/10.1002/prep.202300109
19. Shahzad M, Alhoori H (2022) Public Reaction to Scientific Research via Twitter Sentiment Prediction. *J Data Inform Sci* 7: 97–124. https://doi.org/10.2478/jdis-2022-0003
20. Wang X, Cao Y, Mao B (2020) Spatio-temporal Semantic Analysis of Safety Production Accidents in Grain Depot based on Natural Language Processing. *2020 IEEE/WIC/ACM*

International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) 931–935. https://doi.org/10.1109/WIIAT50758.2020.00142

21. Sun YL, Lian JG, Teng Z, Wei ZY, Tang Y, Yang L, et al. (2024) COVID-19 diagnosis based on swin transformer model with demographic information fusion and enhanced multi-head attention mechanism. *Expert Syst Appl* 243: 122805. https://doi.org/10.1016/j.eswa.2023.122805
22. Zhang H, Yang JW, Dong XB, Lv XG, Jia W, Jin Z, Li XJ (2024) A Video Face Recognition Leveraging Temporal Information Based on Vision Transformer. *Pattern Recognition and Computer Vision: 6th Chinese Conference, PRCV 2023* 29-43. https://doi.org/10.1007/978-981-99-8469-5_3
23. Qian YX (2024) Discriminative Activation of Information Is What You Need in Image Super-Resolution Transformer. *Pattern Recognition and Computer Vision: 6th Chinese Conference, PRCV 2023* 482–493. https://doi.org/10.1007/978-981-99-8552-4_38
24. Chang, MW, Ratinov L, Roth D (2012). Structured learning with constrained conditional models. *Mach Learn* 88: 399–431. https://doi.org/10.1007/s10994-012-5296-5
25. Zhou Y, Fan M, Chen YL, Xiao XQ, Pan XX, Li LH (2024) A transformer model guided by histopathological image information for DCE-MRI-based prediction of response to neoadjuvant chemotherapy in breast cancer. *Medical Imaging 2024: Imaging Informatics for Healthcare, Research, and Applications* 1293114. https://doi.org/10.1117/12.3006656
26. Deng XR, Huang Z, Ma KF, Chen K, Guo J, Qiu WD (2023) GenTC: Generative Transformer via Contrastive Learning for Receipt Information Extraction. *Artificial Neural Networks and Machine Learning – ICANN 2023* 14259. https://doi.org/10.1007/978-3-031-44223-0_32
27. Yuan WL, Chen JX, Chen SF, Feng DW, Hu ZZ, Li P, et al. (2024) Application of Transformer-based reinforcement learning methods in intelligent decision-making: A review. *Front Inform Tech Electr Eng* 25: 763–790. https://doi.org/10.1631/FITEE.2300548.
28. Ochi M, Shiro M, Mori J, Sakata I (2023) Integrating Linguistic and Citation Information with Transformer for Predicting Top-Cited Papers. *Web Information Systems and Technologies. WEBIST 2022* 494: 121–141. https://doi.org/10.1007/978-3-031-43088-6_7
29. Li SY, Dong JW, Chen JY, Gao XZ, Niu SJ (2023) Vision Transformer with Depth Auxiliary Information for Face Anti-spoofing. *Neural Information Processing. ICONIP 2022* 13625: 335–346. https://doi.org/10.1007/978-3-031-30111-7_29
30. Chen JJ, Wang JR, Zheng SL, Liu YJ, Li ZN, Xie SL, et al. (2024) Improving NLOS/LOS Classification Accuracy in Urban Canyon Based on Channel-Independent Patch Transformer with Temporal Information. *Proceedings of the 2024 International Technical Meeting of The Institute of Navigation* 869–882. https://doi.org/10.33012/2024.19507
31. Lin LX, Li YW, Wang HZ (2024) TSMGAN-II: Generative Adversarial Network Based on Two-Stage Mask Transformer and Information Interaction for Speech Enhancement. *Advanced Intelligent Computing Technology and Applications. ICIC 2024* 14865: 174–185. https://doi.org/10.1007/978-981-97-5591-2_15
32. Zhu YP, Huang L, Chen JX, Wang SY, Wan FY, Chen JN (2024) VG-DOCoT: a novel DO-Conv and transformer framework via VAE-GAN technique for EEG emotion recognition. *Front Inform Techn Electr Eng* 25: 1497–1514. https://doi.org/10.1631/FITEE.2300781
33. Liu FY, Zhou ZQ, Men CY, Sun Q, Huang KJ (2023) IFGLT: Information fusion guided lightweight Transformer for image denoising. *J Vis Commun Image Rep* 97: 103994. https://doi.org/10.1016/j.jvcir.2023.103994

34. Xiong WX, Wang P, Sun XC, Wang J (2024) SiET: Spatial information enhanced transformer for multivariate time series anomaly detection. *Knowledge-Based Systems* 296: 111928. https://doi.org/10.1016/j.knosys.2024.111928
35. Abouei E, Pan SY, Hu MZ, Kesarwala AH, Zhou J, Roper J, et al. (2024) Cardiac MRI segmentation using block-partitioned transformer with global-local information integration. *Medical Imaging 2024: Clinical and Biomedical Imaging* 1293021. https://doi.org/10.1117/12.3006929
36. Kesarwani A, Das S, Kisku DR, Dalui M (2024) Dual mode information fusion with pre-trained CNN models and transformer for video-based non-invasive anaemia detection. *Biomed Signal Proces* 88: 105592. https://doi.org/10.1016/j.bspc.2023.105592
37. Zhou LN, Lu ZG, You WK, Fang XF (2023) Reversible data hiding using a transformer predictor and an adaptive embedding strategy. *Front Inform Tech Electr Eng* 24: 1143–1155. https://doi.org/10.1631/FITEE.2300041
38. Ochi M, Shiro M, Mori J, Sakata I (2022) Classification of the Top-cited Literature by Fusing Linguistic and Citation Information with the Transformer Model. *Proceedings of the 18th International Conference on Web Information Systems and Technologies - WEBIST* 286–293. https://doi.org/10.5220/0011542200003318
39. Zhao L, Tian XC, Liu YP (2024) Transformer Based Position Information Enhancement for Medical Image Segmentation. *2024 4th Asia Conference on Information Engineering (ACIE)* 92–96. https://doi.org/10.1109/ACIE61839.2024.00022
40. Hasany SN, Petitjean C, Meriaudeau F (2023) A study of attention information from transformer layers in hybrid medical image segmentation networks. *Medical Imaging 2023: Image Processing* 12464: 389–400. https://doi.org/10.1117/12.2652215
41. Citarella AA, Barbella M, Ciobanu MG, De Marco F, Di Biasi L, Tortora G (2025) Assessing the effectiveness of ROUGE as unbiased metric in Extractive vs. Abstractive summarization techniques. *Journal of Computational Science* 87: 102571. https://doi.org/10.1016/j.jocs.2025.102571
42. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. (2019) PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32: 8024–8035. https://arxiv.org/pdf/1912.01703
43. Liang X, Liu Z, Zhang H (2023) NASTyLinker: NIL-Aware Scalable Transformer-based Entity Linker. In *European Semantic Web Conference* 174–191. https://doi.org/10.1007/978-3-031-33455-9_11
44. Zou X, Zhou X, Zhu Z, Ji L (2019) Novel subgroups of patients with adult-onset diabetes in Chinese and US populations. *The Lancet Diabetes & Endocrinology* 7: 9–11. https://doi.org/10.1016/S2213-8587(18)30316-4

AIMS Press