



UC-TAM: an unsupervised model for fine-grained semantic association construction across large-scale power system standards documents

Weiwei Zheng, Lu Wang, Xiaoqun Yuan, Yanrong Zheng, Wei Xie, Yanan Liu, and Liang Zhao

DOI: <https://doi.org/10.47989/ir31iConf64288>

Abstract

Introduction. Power system standards constitute a vast, complex repository, yet existing methods struggle with fine-grained semantic association (e.g., clause-level) due to coarse-grained analysis and reliance on manual annotation. This study develops an unsupervised, domain-adapted approach to address this gap.

Method. We propose UC-TAM, which integrates a 134,945-term power lexicon with a hierarchical attention framework and unsupervised clustering. Clauses serve as atomic semantic units, where intra- and cross-document relations are captured through multi-head attention and embedded in a shared space, followed by K-means clustering for semantic association discovery.

Analysis. Comparative experiments against a Sentence-BERT + K-means baseline were conducted on subsets of standards, evaluated with Silhouette (SC), Calinski-Harabasz (CH), and Davies-Bouldin (DB) indices.

Results. UC-TAM outperformed the baseline with improvements of 27.7% (SC), 31.3% (CH), and 19.6% (DB). A case study confirmed practical utility, with expert-rated clustering accuracy of 88% versus 75% for the baseline. Performance improved consistently with dataset size, demonstrating scalability and robustness.

Conclusion. UC-TAM highlights both the feasibility and the practical value of unsupervised, domain-aware approaches to standard document analysis, offering a pathway toward more intelligent, automated, and fine-grained management of standards.

Introduction

Standards documents are foundational in the power industry, supporting equipment manufacturing, grid operation, safety management, and regulatory compliance (China Power, 2024). With ongoing digitalisation, the volume of standards has grown rapidly, forming a large repository that spans multiple technical sub-domains (Huang & Lin, 2023). In this context, efficient management and utilisation of standard documents have become increasingly important. Existing work has investigated metadata extraction for domain-specific documents, including structured storage of metadata and terminology (Zheng et al., 2022) and large language model-assisted processing (Chen et al., 2025), which provide the basis for standardised data organisation. At the same time, uncovering semantic relationships among standards is gaining importance. Fine-grained clause-level semantic associations can support information retrieval, collaborative drafting, and knowledge reuse. For example, identifying similar clauses helps engineers locate related provisions across documents and maintain consistency when drafting new regulations.

However, research on semantic association for standards remains limited. Most existing studies focus on document-level or topic-level modeling (Wang et al., 2017), yielding coarse semantic signals with limited interpretability (Tian et al., 2023). Other approaches rely heavily on expert annotation, which is costly and difficult to scale as new standards continue to emerge (Paulheim, 2016). As a result, systematic exploration of fine-grained semantic associations across large standard corpora remains insufficient.

In this paper, we propose UC-TAM, an unsupervised clustering model based on a *terminology-enhanced hierarchical attention mechanism*. Domain-specific terminology is incorporated into attention computation to improve clause-level semantic representation, while hierarchical multi-head attention captures structural similarities across documents. K-means clustering is then applied to group semantically related clauses. Experiments on real power system standards show that UC-TAM achieves superior performance over a baseline in both quantitative metrics and qualitative semantic coherence.

The main contributions of this work are as follows.

1. We propose a novel unsupervised framework for fine-grained semantic association of power standards, incorporating a terminology-enhanced hierarchical attention mechanism tailored for domain specificity.
2. We empirically validate the effectiveness and stability of the proposed model through extensive experiments, showing its advantage both in clustering metrics and in practical semantic accuracy at the clause level.

Methods

Problem definition

Definition1: Semantic association

Let $D = \{d_1, d_2, \dots, d_N\}$, denote a collection of N standard documents, where each document d_i consists of a set of clauses $d_i = \{s_{i1}, s_{i2}, \dots, s_{iM_i}\}$. Each clause s_{ij} is represented as a sentence-level semantic vector $\mathbf{h}_{ij} \in R^d$. We define the semantic association between two clauses s_{ij} and s_{kl} as their similarity in the semantic embedding space, measured by a distance function $\delta(\cdot, \cdot)$, e.g., cosine similarity. Formally,

$$\text{Assoc}(s_{ij}, s_{kl}) = f(\mathbf{h}_{ij}, \mathbf{h}_{kl}), \quad f: R^d \times R^d \rightarrow R.$$

A semantic association is considered significant if $\text{Assoc}(s_{ij}, s_{kl}) \geq \tau$, where τ is a predefined similarity threshold.

Definition2: Research problem

Given a corpus of power standard documents D , the research problem is to design a function F that maps all clauses into semantically meaningful clusters without supervised labels,

$$F: \{s_{ij}\}_{i=1, j=1}^{N, M_i} \rightarrow \{C_1, C_2, \dots, C_K\},$$

where C_k denotes the k -th cluster of semantically related clauses, and K is determined automatically by the clustering algorithm. The goal is to ensure that intra-cluster semantic similarity is maximised while inter-cluster similarity is minimised.

Framework of UC-TAM

To tackle the above problem, we propose **UC-TAM**, an **unsupervised clustering model based on a terminology-enhanced hierarchical attention mechanism**. The schematic diagram of the proposed framework is shown in **Figure 1**.

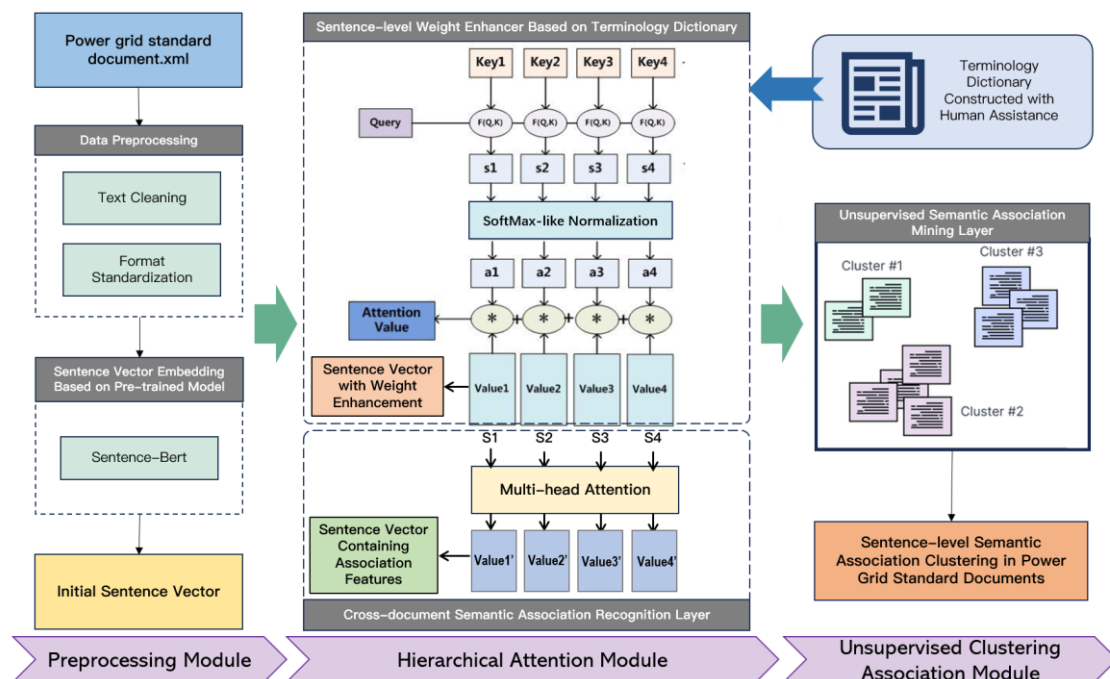


Figure 1. Technical framework of the proposed model.

Terminology-enhanced hierarchical attention network

The design goal of this module is to strengthen the embedding representation of power-domain clauses by automatically identifying and weighting domain-specific terminology. Unlike generic embeddings, this approach ensures that the model captures both semantic nuance and structural association at the clause level.

Power terminology lexicon construction

To further adapt the general embedding model to the professional characteristics of power system standards, we construct a domain-specific terminology lexicon that explicitly incorporates technical vocabulary into the attention mechanism.

The lexicon was primarily derived from two authoritative sources, the national standard GB/T 2900.1—2008 ‘Electrotechnical Terminology—Fundamental Terms’ and the *China Electric Power Encyclopedia* (中国电力百科网), which consolidates standard terminology from industry, national, and enterprise standards. The resulting resource contains a total of 134,945 domain-specific entries. During the curation process, we emphasised terms with strong domain specificity while

excluding high-frequency units that also appear frequently in general corpora. Particular attention was given to composite expressions that embody typical operational semantics in the power industry, such as ‘differential protection setting value (差动保护整定值)’ and ‘synchronous closing condition (同期合闸条件)’.

Sentence-level terminology injection

Given a sentence s_{ij} consisting of tokens $\{w_1, w_2, \dots, w_L\}$, each token is initially embedded by BERT into $x_t \in R^d$. We introduce a terminology indicator function $\gamma(w_t)$, defined as

$$\gamma(w_t) = \begin{cases} \alpha > 1, & w_t \in \text{Lexicon} \\ 1, & \text{otherwise,} \end{cases}$$

where α is a hyperparameter controlling the relative importance of terminology tokens. The weighted embedding is thus defined as

$$\tilde{x}_t = \gamma(w_t) \cdot x_t.$$

The sentence embedding is then obtained by an attention-weighted sum,

$$h_{\{ij\}} = \sum_{t=1}^L \beta_t \tilde{x}_t, \quad \beta_t = \frac{\exp(u^T \tanh(W \tilde{x}_t))}{\sum_{k=1}^L \exp(u^T \tanh(W \tilde{x}_k))}$$

where β_t is the attention weight, W is a learnable projection matrix, and u is the attention vector.

In our implementation, clause representations are obtained using a pre-trained sentence encoder (*paraphrase-mpnet-base-v2*). The encoder is used as a fixed feature extractor throughout all experiments, and no domain-specific fine-tuning is performed. Sentence embeddings are derived from the last hidden layer by selecting the first token representation.

Cross-document interaction embedding

To capture potential semantic alignment across documents, we employ a multi-head attention mechanism (Vaswani, et al., 2017). Given sentence embeddings $\{h_{ij}\}$ from multiple documents, the cross-document interaction is modeled as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where $Q = HW_Q, K = HW_K, V = HW_V$, and H is the matrix of all sentence embeddings. Multi-head attention is then computed as

$$\text{MultiHead}(H) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W_O$$

with

$$\text{head}_i = \text{Attention}(HW_Q^i, HW_K^i, HW_V^i)$$

The resulting interaction-enhanced embedding z_{ij} is used as the final clause representation for clustering,

$$z_{ij} = \text{MultiHead}(H)[ij]$$

Unsupervised clustering for semantic association

After obtaining sentence-level semantic embeddings enhanced by the terminology-aware attention mechanism, unsupervised K-means algorithm (Ikotun, et al., 2023; Likas, et al., 2003) is applied to discover semantic associations across documents through clustering.

Clustering method

Let $z_{ij} \in R^d$ be the interaction-enhanced embedding of clause s_{ij} produced in Section 2.3. Let λ_{ij} be the sentence-level terminology weight (as defined below),

$$\lambda_{ij} = \frac{1}{L_{ij}} \sum_{t=1}^{L_{ij}} \gamma(w_t), \quad \gamma(w_t) = \begin{cases} \alpha > 1, & w_t \in \text{Lexicon} \\ 1, & \text{otherwise,} \end{cases}$$

and define the terminology-rescaled embedding

$$\hat{z}_{ij} = \lambda_{ij} \cdot z_{ij}$$

We cluster $\{\hat{z}_{ij}\}$ with a **terminology-weighted** K-means (K-means++ initialisation, L_2 distance). For a candidate K and clusters $\{C_1, \dots, C_K\}$, the objective is

$$\min_{\{C_k\}, \{\mu_k\}} \sum_{k=1}^K \sum_{s_{ij} \in C_K} \lambda_{ij} \|\hat{z}_{ij} - \mu_k\|^2$$

where the weighted centroid is

$$\mu_k = \frac{\sum_{s_{ij} \in C_K} \lambda_{ij} \hat{z}_{ij}}{\sum_{s_{ij} \in C_K} \lambda_{ij}}.$$

Unsupervised tuning of parameters

To automatically determine the number of clusters K and the terminology weighting parameter α , we adopt a two-stage unsupervised search strategy. First, we evaluate candidate values of K within a pre-defined range (e.g., 2–50) using internal clustering validity metrics. The optimal K is selected by maximising the Silhouette coefficient and Calinski–Harabasz index while minimising the Davies–Bouldin index. Second, for each candidate K , we tune the terminology enhancement factor α to balance the contribution of domain-specific terminology and general semantic context. A grid search with repeated trials is conducted, and the pair (K, α) that achieves the best average performance across multiple runs is selected as the final setting. This procedure avoids reliance on labelled data and ensures robust clustering results in large-scale unsupervised scenarios. The parameter tuning process is fully unsupervised and does not rely on any labelled data or human intervention. All candidate configurations are evaluated using internal clustering metrics on the same representation space, and the selected parameters are applied consistently across all datasets.

Cluster evaluation metrics

We employ three widely used unsupervised clustering evaluation metrics:

1. **Silhouette coefficient (SC).** Measures cohesion and separation of clusters. A higher value (i.e., approaching 1) indicates that samples are well matched to their own cluster and poorly matched to neighboring clusters (Van Wyk & Van Vuuren, 2021).
2. **Calinski–Harabasz index (CH).** Evaluates the ratio of between-cluster variance to within-cluster variance. Larger values imply that clusters are more compact and better separated (Mokashi & Lepakshi, 2024).
3. **Davies–Bouldin index (DB).** Quantifies average similarity between each cluster and its most similar one. Lower values indicate better cluster separability (Liu, 2022).

These three metrics together provide a comprehensive evaluation of clustering quality, covering cohesion, separation, and overall structural discrimination.

Experiments

Setup

Dataset

We randomly collected a subset of 61 power system standards, covering diverse authoritative levels (i.e., national, industry, group, enterprise, and local standards), and multiple categories (i.e., terminology, tests, specifications, regulations, and guidelines). All documents were converted into XML format, and sentence-level clauses were extracted as semantic units. For experiments, we sampled subsets of 3, 5, and 10 documents from a 20-document pool and repeated each sampling 20–30 times to reduce randomness.

Compared baseline

We compare our UC-TAM model against a baseline that directly applies Sentence-BERT embeddings without domain adaptation.

All experiments are conducted in PyTorch with HuggingFace Transformers. The pre-trained encoder (*paraphrase-mpnet-base-v2*) is kept frozen. The attention modules are randomly initialised and trained for 10 epochs with batch size 16 using Adam ($lr = 1e-4$).

General performance

Parameter sensitivity experiments

To analyse hyperparameter effects, we perform a grid search over the cluster number K ($\in [5, 30]$) and the terminology weighting factor α ($\in [1.0, 3.0]$). For each (K, α) pair, we run the full pipeline 20 times with different seeds and report the average Silhouette, Calinski–Harabasz, and Davies–Bouldin scores. Table 1 partially illustrates the parameter sensitivity as summarised visualisation. The optimal parameters on our validation grid were set to be $K^*=30, \alpha^*=1.6$, with average Silhouette = 0.178 on the validation runs. The setting works for the following full-scale comparative experiments. For small-scale comparability tests reported elsewhere in this paper, we still use fixed $K=10$ when explicitly stated; the above K^* refers to the global parameter search on a larger pooled corpus.

K	α	Silhouette (mean)
5	1.0	0.152
5	1.6	0.155
10	1.6	0.16
15	1.6	0.165
20	1.6	0.169
25	1.6	0.173
30	1.6	0.178

Table 1. Excerpt of parameter sensitivity results (20 runs average).

The trend of the Silhouette coefficient changing with the number of clusters (K) under different terminology weighting factors (α) is visually presented in Figure 2. Moderate weighting around 1.5–1.8 yields the best trade-off. A too-small α under-utilises domain terminology and a too-large α over-emphasises terminology tokens (which may fragment clusters around rare terms).

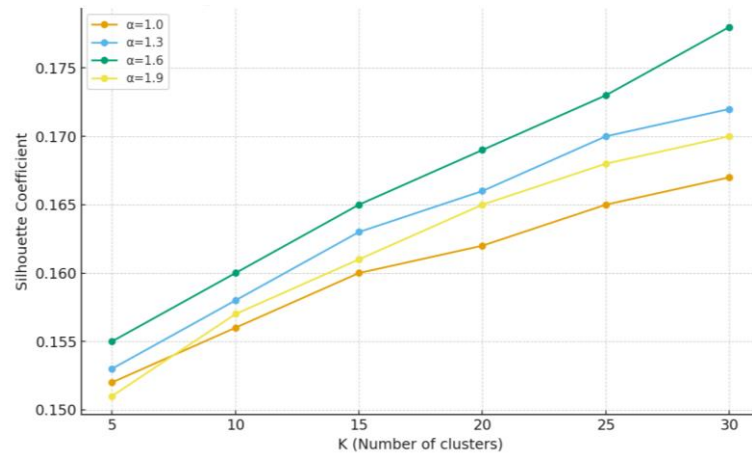


Figure 2. Parameter sensitivity curves (Silhouette vs K).

Preliminary comparison on sampled subsets

To evaluate model effectiveness under controlled settings, we conducted small-scale comparisons using three subsets of 3, 5, and 10 documents sampled from 20 candidate standards. For each subset, both models were run repeatedly with the number of clusters fixed at 10 for comparability (consistent with the parameter sensitivity configuration). Numerical results are reported in Table 2 and clustering visualisations are shown in Figure 3.

Corpus Size	Avg. Clause Units	Metric	UC-TAM	Baseline
3 docs	185±12	SC	0.145	0.120
		CH	62.5	51.0
		DB	2.45	2.78
5 docs	312±18	SC	0.158	0.132
		CH	74.2	60.3
		DB	2.20	2.55
10 docs	628±25	SC	0.172	0.140
		CH	88.6	70.1
		DB	1.98	2.42

Table 2. Small-scale comparison of proposed vs. baseline (K = 10).

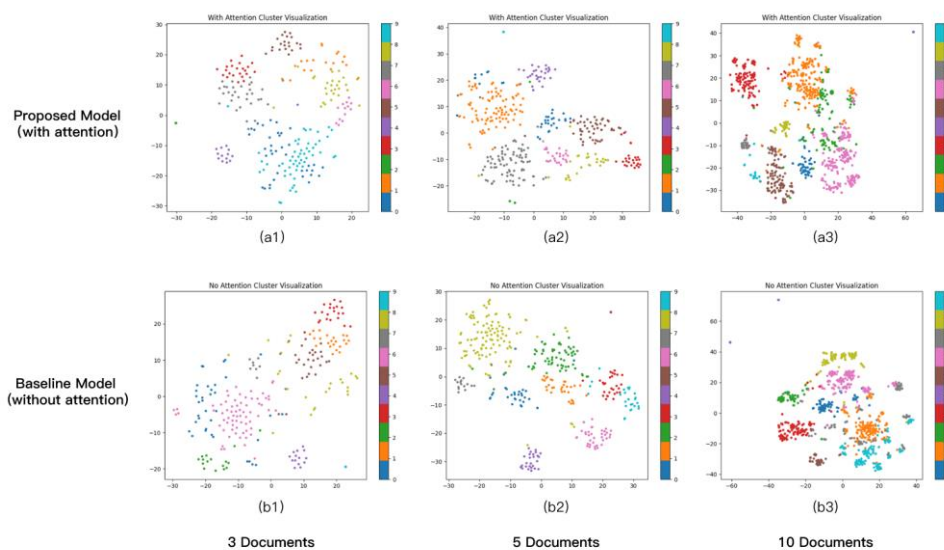


Figure 3. Semantic association visualisation of clustering

The results show that UC-TAM consistently surpasses the baseline across corpus sizes, yielding 20.8–22.9% gains in Silhouette Coefficient, 22.5–26.3% in Calinski–Harabasz index, and 13.7–18.2% reductions in Davies–Bouldin index. The performance gap increases with corpus size, indicating clearer boundaries, better cluster separations, and improved scalability for large power standard corpora.

Full-corpus evaluation with optimal parameters

To demonstrate the model's scalability and effectiveness on a comprehensive, real-world scale, we conducted the final evaluation on our entire collection of 61 power system standards. Using the optimal configuration ($K^*=30$, $\alpha^*=1.6$), we performed this large-scale comparison. To obtain robust statistics, each full experiment was repeated 30 times with different random seeds (for both subset sampling and K-means initialisation). Figure 4 illustrates the results. Compared with the baseline, UC-TAM achieved significantly higher clustering performance across all metrics. In average, the Silhouette Coefficient improved from 0.141 ± 0.013 to 0.180 ± 0.012 , representing a 27.7% increase ($t = 14.03$, $p < 0.001$). The Calinski–Harabasz index rose from 79.2 ± 7.0 to 104.0 ± 8.0 , a 31.3% improvement ($t = 9.78$, $p < 0.001$). Conversely, the Davies–Bouldin index dropped from 2.39 ± 0.14 to 1.92 ± 0.12 , corresponding to a 19.6% reduction ($t = -12.68$, $p < 0.001$). These results confirm the significant advantage and robustness of our model in generating clearer cluster boundaries and sharper separations at a large scale.

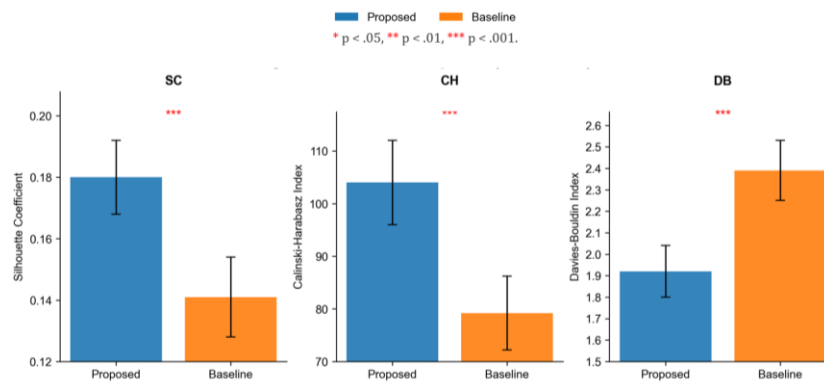


Figure 4. Performance of large-scale comparisons between UC-TAM and baseline (30 runs, paired t-tests).

Case study

To qualitatively assess the semantic coherence of the generated clusters, we conducted a case study based on the results of the 10-document experiment. Table 3 reports the expert evaluation results. For each model, three representative clusters (denoted as C1, C2, and C3) were randomly selected. Five experts with professional backgrounds in power system standardisation, electrical engineering research, and technical standard drafting were invited to independently assess the semantic consistency of clauses within each cluster.

The experts were asked to judge whether the clauses grouped in each cluster conveyed coherent and conceptually aligned semantic meanings, and to report the proportion of clauses they considered semantically consistent. The evaluation was conducted independently without access to model identities. On average, UC-TAM achieved an expert-evaluated semantic consistency score of 88.3%, outperforming the baseline by 17.73%. These results indicate that incorporating domain-specific terminology and cross-document attention contributes to improved semantic coherence in clustered standard clauses.

Cluster ID	Representative Clauses (summary)	Expert Accuracy (UC-TAM)	Expert Accuracy (Baseline)
C1	Clauses on equipment safety thresholds and overload protection.	88%	78%
C2	Clauses on synchronisation, parallel operation, and grid stability.	92%	80%
C3	Clauses on environmental impact limits and compliance guidelines.	85%	67%

Table 3. Representative clause summaries and expert accuracy comparison in case study.

Discussion

This study investigates fine-grained semantic association mining in power system standards, a domain with large document volume, technical complexity, and strong terminology dependence. We define sentence-level semantic association and propose an unsupervised clustering framework with terminology-enhanced hierarchical attention. Incorporating a domain terminology lexicon into attention computation improves the capture of field-specific semantics, while unsupervised clustering enables scalable association discovery without manual labeling.

Compared with prior work focusing on document- or topic-level associations or relying on expert annotation, our approach contributes in three aspects. First, it formalises clause-level semantic

association for standards and demonstrates its practicality for downstream tasks. Second, it integrates power domain terminology into a two-stage attention design combined with unsupervised clustering, improving representational accuracy in specialised technical contexts while preserving scalability. Third, it supports digital standardisation scenarios such as semantic search, collaborative drafting, automatic keyword extraction, and preprocessing for vertical-domain LLMs.

Experimental results show that the framework yields semantically coherent clusters aligned with domain knowledge and achieves favorable internal clustering scores. Qualitative expert evaluation further indicates that terminology enhances semantic alignment across clauses. These findings support the feasibility of unsupervised domain-aware analysis and suggest utility for automated standard structuring and management.

Limitations remain that the current batch-mode clustering pipeline requires re-embedding and re-clustering when new standards are introduced, reducing adaptability in dynamic ecosystems. Future work will investigate incremental update mechanisms for real-time association refresh, as well as expanded expert evaluation and benchmark resources as data access improves within the standardisation domain.

Acknowledgements

This work was supported by State Grid Corporation of China (Comprehensive Construction Technology of Standard Text Resources and Key Elements, 5700-202366831A-4-2-KJ), and Key Laboratory of Semantic Publishing and Knowledge Service of the National Press and Publication Administration (Wuhan University).

About the authors

Weiwei Zheng is a master's Student in School of Information Management, Wuhan University, China. Her research interests are cultural big data and data mining. She is the co-first author and can be contacted at 2023201040074@whu.edu.cn

Lu Wang is a master's Student in School of Information Management, Wuhan University, China. Her research interests are cultural big data and data mining. She is the co-first author and can be contacted at luwang23@whu.edu.cn

Xiaoqun Yuan is Associate Professor in School of Information Management, Wuhan University, China. His research interests are information resource management and knowledge services and can be contacted at yuan20030308@whu.edu.cn

Yanrong Zheng is a Senior Editor at Yingda Media Investment Group Co., Ltd. Her research interest is digitalisation of electric power standards, and she can be contacted at 80653805@qq.com

Wei Xie works at State Grid Fujian Electric Power Research Institute, Fuzhou, China. His research interest is artificial intelligence in electric power and can be contacted at xiewei3896@163.com

Yanan Liu is an Editor at Yingda Media Investment Group Co., Ltd. Her research interest is digitalisation of electric power standards, and she can be contacted at 743113090@qq.com

Liang Zhao is an Associate Professor in School of Information Management, Wuhan University, China. Her research interests are data mining and knowledge services. She is the corresponding author of the paper and can be contacted at liangzhao@whu.edu.cn

References

- Chen, G., Xie, W., Liu, Y., Yuan, X., & Zhao, L. (2025). Systematically modeling and extracting bibliographic metadata of power grid standard documents with LLMs. *Information Research an International Electronic Journal*, 30(iConf), 654–665. <https://doi.org/10.47989/ir30iConf47233>
- China Electric Power Encyclopedia. (2025). State Grid Corporation of China & Yingda Media Investment Group Co.,Ltd. <https://www.ceppedu.com/home/search-term.html>
- China Power. (2024, July 19). State Grid lays a solid foundation for the construction of a new type of power system.
- Huang, C., & Lin, B. (2023). Promoting decarbonisation in the power sector: How important is digital transformation?. *Energy Policy*, 182, 113735. <https://doi.org/10.1016/j.rser.2018.05.068>
- Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, 178–210. <https://doi.org/10.1016/j.ins.2022.11.139>
- Liu, G. (2022). A New Index for Clustering Evaluation Based on Density Estimation. *arXiv:2207.01294*. <https://arxiv.org/abs/2207.01294>
- Mokashi, R., & Lepakshi, V. A. (2024). Enhancing Pattern Classification Accuracy Through Customer Segmentation Using Machine Learning Algorithms. *SN Computer Science*, 5(7), 948–17. <https://dblp.org/rec/journals/snscs/MokashiL24a.html>
- Paulheim, H. (2016). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, 8(3), 489–508. <https://doi.org/10.3233/SW-160218>
- Tian, D., Li, M., Shen, Y., & Han, S. (2023). Intelligent mining of safety hazard information from construction documents using semantic similarity and information entropy. *Engineering Applications of Artificial Intelligence*, 119, 105742. <https://doi.org/10.1016/j.engappai.2022.105742>
- Van Wyk, J. J., & Van Vuuren, J. H. (2021). Unsupervised Fine-tuning of Speaker Diarisation Pipelines using Silhouette Coefficients. *North-West University Technical Report*. <https://ieeexplore.ieee.org/abstract/document/10735414>
- Vaswani, A., Shazeer, N., Parmar, et al. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. <https://doi.org/10.48550/arXiv.1706.03762>
- Wang, Y. et al. (2017). Topic Model Based Text Similarity Measure for Chinese Judgment Document. In: Zou, B., Han, Q., Sun, G., Jing, W., Peng, X., Lu, Z. (eds) *Data Science. ICPCSEE 2017. Communications in Computer and Information Science*, vol 728. Springer, Singapore. https://doi.org/10.1007/978-981-10-6388-6_4
- Zheng, L., Pan, J., & Zhang, K. (2022). Power Data Integration Method Based on Database-table Metadata Semantic. In *Journal of Physics: Conference Series* (Vol. 2179, No. 1, p. 012028). IOP Publishing. <https://doi.org/10.1088/1742-6596/2179/1/012028>

© [CC-BY-NC 4.0](#) The Author(s). For more information, see our [Open Access Policy](#).