Human Factors in Training, Education, and Learning Sciences, Vol. 231, 2026, 181–185
https://doi.org/10.54941/ahfe1008009

Semantic Structure and Importance Extraction from Sequential Conversational Data via Dimensional Reduction

Takeshi Matsuda[1,2] and Michio Sonoda[2]

[1]Faculty of Information Sciences, Hannan Universoyt, Matsubara Osaka 580-0032, Japan
[2]National Institute of Information and Communications Technology (NICT), Tokyo, Japan

ABSTRACT

The analysis of spoken data from panel discussions, policy dialogues, and educational meetings has gained increasing importance in both academic research and professional practice. However, conventional approaches to Japanese conversation analysis have relied heavily on keyword matching or surfacelevel text similarity, making it difficult to capture deeper semantic relationships, topic transitions, and latent discourse structures. In addition, Japanese natural language processing pipelines often rely on environment-sensitive morphological analyzers, which hinder reproducibility and large-scale processing. To address these limitations, this study proposes a robust and semantically enriched framework for conversation understanding based on a composite distributed representation.

The proposed method integrates three layers of linguistic information: (1) contextual sentence embeddings generated by a multilingual transformer model, (2) word embeddings obtained from fastText, and (3) cooccurrence vectors that capture lexical association patterns within the conversation. Sudachi is employed for Japanese text preprocessing to ensure stable and reproducible morphological analysis. By combining these components into a unified composite vector, the framework simultaneously represents global sentencelevel meaning and local lexical relationships. Using this representation, a directed graph is constructed that incorporates both temporal adjacency and semantic proximity between utterances, enabling the visualization of key conversational connections. To evaluate the effectiveness of the composite representation, dimensionalityreduction algorithms are applied to examine whether semantically similar utterances naturally form coherent clusters in lowdimensional space. The resulting clusters are assessed for consistency and interpretability, demonstrating that the proposed representation successfully captures meaningful conversational structure.

Keywords: Conversational semantics, Distributed representations, Dimensionality reduction (MAPE), Nonlinear embedding, Semantic clustering

INTRODUCTION

Understanding how meaning emerges and shifts within natural conversation is a central challenge in language and education research. Spoken dialogue is characterized by short, context-dependent utterances, rapid topic transitions,

Received May 31, 2026; Revised June 16, 2026; Accepted June 24, 2026; Available online July 20, 2026

and heterogeneous semantic content, making it difficult for conventional analytical methods to capture its latent structure. Traditional approaches often rely on surface-level lexical features or manually defined categories, which struggle to represent the layered, distributed nature of conversational meaning. As a result, the semantic organization of dialogue—how utterances cluster, diverge, or form thematic regions—remains insufficiently understood [Yerin Hwang, Xiang Li].

To address this gap, the present study integrates three complementary sources of semantic information: contextual embeddings derived from large language models, lexical features reflecting word-level properties, and co-occurrence patterns capturing distributional relationships across the dataset. These components are combined into a unified 384-dimensional distributed representation for each utterance, enabling a richer modeling of latent semantic structure than is possible with conventional techniques.

We then apply dimensionality reduction and clustering to examine whether coherent semantic regions can be identified within this high-dimensional space. In particular, we compare PCA, a widely used linear method, with MAPE [T. Matsuda], a nonlinear algorithm designed to preserve both local and global similarity structures through a weighted combination of cosine similarity and Euclidean distance. While PCA yields only moderate cluster separation, MAPE produces a clear and interpretable three-cluster structure, with a substantially higher silhouette score. These results demonstrate that nonlinear modeling provides a more faithful representation of conversational semantics, revealing distinct regions corresponding to lightweight backchannels, interpersonal and collaborative discourse, and institutional or administrative topics.

Overall, this study shows that integrating multiple semantic signals and applying nonlinear dimensionality reduction enables a more precise understanding of how meaning is organized in natural conversation. The findings highlight the value of distributed representations and structure-preserving embeddings for analyzing complex, context-rich linguistic data.

Proposed Method

Our proposed method constructs a composite distributed representation of each utterance by integrating three complementary sources of semantic information: contextual embeddings, lexical embeddings, and co-occurrence vectors. First, we obtain contextual embeddings using a Sentence-BERT model, capturing utterance-level meaning shaped by surrounding linguistic context. Second, we compute averaged fastText word embeddings to incorporate lexical-level semantic properties. Third, we derive co-occurrence vectors from the entire dataset to represent distributional relationships among words and utterances. These three components are concatenated to form a unified 384-dimensional representation.

To analyze the latent semantic structure of the conversation, we apply MAPE, a nonlinear dimensionality reduction algorithm that preserves both

local and global similarity structures through a weighted combination of cosine similarity and Euclidean distance. MAPE models pairwise attraction using a mixture of distance-based distributions and optimizes an attraction–repulsion loss to obtain low-dimensional coordinates. The resulting 3D embedding is then subjected to clustering analysis to identify coherent semantic regions. This integrated approach enables a more faithful representation of conversational meaning than linear methods such as PCA.

RESULTS

Distributed Representations and Dimensionality Reduction

Each utterance was represented as a 384-dimensional composite vector integrating contextual embeddings, lexical features, and co-occurrence information. To examine the latent semantic structure, we applied two dimensionality reduction methods: PCA and MAPE. While PCA captured only moderate global variance, MAPE successfully preserved both local and global similarity structures through its nonlinear attraction–repulsion optimization.

MAPE-Based Embedding and Cluster Structure

MAPE reduced the 384-dimensional vectors to a three-dimensional embedding that revealed a clear and interpretable cluster structure. Clustering analysis identified three distinct semantic regions, achieving a silhouette score of 0.6702, indicating strong separation and coherent internal organization.

Cluster 0: Short replies, backchannels, transitional utterances
Cluster 1: Human relationships, collaboration, early-childhood–elementary transition
Cluster 2: Institutional and administrative discourse (MEXT, standards, programs)

The 3D scatter plot showed that these clusters were well separated, reflecting meaningful semantic distinctions within the conversational data.

PCA-Based Embedding and Weaker Separation

In contrast, PCA produced substantially weaker cluster separation. Silhouette scores for k = 2–14 indicated that k = 2 was optimal, but the highest score was only 0.3675, suggesting a moderate and less coherent structure.

PCA Cluster 0 contained a diffuse mixture of impressionistic and context-dependent utterances.

PCA Cluster 1 grouped institutional vocabulary but mixed in relational and collaborative elements, reflecting PCA's coarse semantic resolution.

Overall, PCA failed to isolate lightweight utterances or administrative discourse as clearly as MAPE.

Comparison of Methods

Across all analyses, MAPE consistently outperformed PCA in revealing interpretable semantic regions. The nonlinear preservation of similarity structures enabled MAPE to distinguish subtle conversational functions—such as backchannels, interpersonal discourse, and administrative terminology—that PCA compressed into broader, less coherent clusters.

CONCLUSION

This study examined how latent semantic structure emerges within natural conversation by integrating multiple sources of semantic information and applying nonlinear dimensionality reduction. By combining contextual embeddings, lexical features, and co-occurrence patterns into a unified distributed representation, we were able to model conversational meaning beyond what surface-level features can capture. The comparison between PCA and MAPE demonstrated that nonlinear, structure-preserving methods provide a substantially clearer view of semantic organization. MAPE produced a well-defined three-cluster structure with strong separation, revealing distinct regions corresponding to lightweight backchannels, interpersonal and collaborative discourse, and institutional or administrative topics. In contrast, PCA yielded only moderate separation and mixed semantically heterogeneous utterances into broader clusters.

These findings highlight the importance of nonlinear modeling for analyzing context-rich, heterogeneous conversational data. The proposed approach offers a more faithful representation of how meaning is distributed across utterances and provides a foundation for future work on dialogue understanding, educational discourse analysis, and automated semantic assessment. Overall, the results demonstrate that integrating diverse semantic signals with nonlinear dimensionality reduction is an effective strategy for uncovering the underlying structure of natural conversation.

Table 1: Summary of results across all analyses.

Perspective	MAPE	PCA
Optimal number of clusters	k = 3 (silhouette = 0.67)	k = 2 (silhouette = 0.36)
Cluster clarity	Very clear (high separation)	Moderate (clusters mix easily)
Semantic structure	Three naturally emerging semantic regions	Two coarse, broad regions
Consistency of representative utterances	High (semantically coherent)	Low-medium (heterogeneous)
Ability to extract latent conversational structure	Very strong	Limited

REFERENCES

T. Matsuda (2025). https://pypi.org/project/mape/0.1.2/

T. Matsuda and M. Sonoda (2025). Embedding Method for Structural Preservation via Pairwise Attractiveness; Proceedings of the 57th ISCIE International Symposium on Stochastic Systems Theory and Its Applications.

Xiang Li, et al. (2025). A Hierarchical Framework for Dialogue Topic Segmentation via Global Topic Shifts and Local Entity Coherence, Advanced Intelligent Computing Technology and Applications, pp. 77–88

Yerin Hwang, et al. (2024). MP2D: An Automated Topic Shift Dialogue Generation Framework Leveraging Knowledge Graphs, Proceedings of the 2024 Conference on Empirical Methods in Natural Language, p. 17682–17702