Automated Framework for Diachronic Chinese WordNet Construction with Diachronic Contextual Embedding

Anonymous submission

Abstract

While the study in diachronic semantic change have advanced with recent computational developments, structured lexical resources that reflect semantic evolution remain scarce for many languages. This study presents a robust, automated methodology for constructing a Diachronic Chinese WordNet (D-CWN). Our pipeline operates on historical corpora partitioned into eight dynastic periods (戰國 Warring States, 秦漢 Qin–Han, 魏晉 Wei–Jin, 隋唐 Sui–Tang, 宋元 Song–Yuan, 明 Ming, 清 Qing, 民國 Republi- can period). It employs GuwenBERT, a pre-trained language model for Classical Chinese, to generate contextualized embeddings from punctuated texts using sentence-level contexts. Within each period, K-means clustering discovers distinct word senses, with optimal cluster numbers determined by the Elbow method, followed by PCA for visualization. These senses are then aligned across consecutive dynasties using N×M average pairwise cosine distance between all embedding pairs, allowing us to classify evolutionary patterns through quartile-based thresholds. A pilot study on the character 手 validates the methodology, demonstrating its ability to robustly quantify semantic continuity and shift across more than two millennia of textual data. Our work establishes a scalable framework for creating the first large-scale, data-driven diachronic lexical resource for Chinese, bridging the gap between statistical semantic analysis and structured lexicography.

Keywords: Diachronic WordNet, Chinese NLP, Semantic Change, Historical Linguistics

1. Introduction

Lexical-semantic change is fundamental to language evolution. Despite diachronic WordNets for several Indo-European languages, Chinese still lacks a comprehensive, computationally derived resource. Existing Chinese WordNets (CWN, COW, MCW) are strictly synchronic and centered on Modern Mandarin, thus missing the polysemy, specialization, and drift accumulated over three millennia of usage. Literary Chinese, attested continuously from the pre-Qin period onward, shows rich lexico-semantic shifts absent from current resources. This study proposes a method to automatically construct a diachronic CWN (D-CWN) partitioned by historical periods, each modeled as a structured semantic space. The resource enables (1) quantitative measurement of semantic change for any lexeme across dynasties and (2) automatic discovery of novel senses, facilitating the tracing of conceptual evolution in historical context. Our approach adapts recent computational techniques in diachronic semantics to Chinese-specific challenges—logographic script, lack of word boundaries, and vast historical corpora.

2. Related Work

2.1. Chinese WordNet Development

Princeton WordNet (PWN) defined the synset–relation framework that underpins modern lexical resources (Fellbaum, 1998; Miller, 1994). Building on PWN, Chinese WordNet (CWN) by Academia Sinica and National Taiwan University pioneered Chinese lexical resources and enabled Sinica BOW via bilingual PWN–CWN alignment (Huang et al., 2004; Lee et al., 2009). Subsequent efforts—Chinese Open WordNet (COW) within Open Multilingual WordNet and Multi-Fusion Chinese WordNet (MCW)—expanded coverage (Bond and Foster, 2013; Wang and Bond, 2013; Li et al., 2020). These resources provide manually curated synsets for contemporary Mandarin and support many Chinese NLP applications (Huang et al., 2010), yet they remain fundamentally synchronic, offering no representation of historical sense evolution or temporally varying conceptual relations.

2.2. Computational Semantic Change Detection

Semantic change is commonly modeled in distributional spaces built from temporally partitioned corpora. A standard approach trains separate embeddings per slice and quantifies movement via cosine distance or neighborhood shifts (Hamilton et al., 2016). Dynamic models encode time directly to obviate post-hoc alignment (Rudolph and Blei, 2018), while other methods enforce a shared coordinate system across periods (Carlo et al., 2019). Surveys highlight persistent issues—domain drift, corpus comparability, and evaluation reliability—motivating stronger protocols (Tahmasebi et al., 2019; Kutuzov et al., 2018). Shared tasks (SemEval-2020 Task 1, LSCDiscovery, DiaCR-Ita) help standardize datasets and metrics across languages (Schlechtweg et al., 2020; Zamora-Reina et al., 2022; Basile et al., 2020).

2.3. Word Sense Induction and Discovery

Unsupervised sense discovery through context clustering dates to Schütze (Schütze, 1998). Modern approaches use contextualized embeddings, with BERT-based substitution methods improving cluster quality (Amrami and Goldberg, 2019). Determining optimal cluster numbers remains challenging, with silhouette coefficients, gap statistics, and elbow methods providing complementary perspectives. Sense-level representations linked to lexical resources offer ways to anchor clusters to existing inventories and compare diachronic prototypes directly (Rothe and Schütze, 2015).

2.4. Classical Chinese Language Resources and Models

Major diachronic corpora include the *Chinese Text Project* (ctext)(Sturgeon, 2011), a curated digital library of Classical Chinese covering literary, philosophical, and historiographic registers. It offers broad temporal and domain coverage, with bibliographic metadata that supports reliable periodization. Pre-trained models for Classical Chinese include *AnchiBERT* (Tian et al., 2021) and *GuwenBERT*(Ethan-yt, 2020), which address archaic lexicon and orthographic variation. *C3Bench* supplies evaluation benchmarks for Classical Chinese understanding (Cao et al., 2024). Recent shared tasks (Li et al., 2022, 2024) document steady progress while underscoring remaining challenges.

3. Proposed D-CWN Methodology

Our methodology comprises three stages: (1) **within-dynasty sense discovery** through K-means clustering followed by PCA visualization, (2) **cross-dynasty sense alignment** using average pairwise distance to identify evolutionary relationships, and (3) **temporal-semantic analysis** to quantify and classify evolutionary patterns. This pipeline operates automatically on punctuated historical Chinese texts, requiring minimal human intervention.

3.1. Within-Dynasty Sense Discovery

For each temporal slice T_i (戰國 *Warring States,* 秦漢 *Qin–Han,* 魏晉 *Wei–Jin,* 隋唐 *Sui–Tang,* 宋元 *Song–Yuan,* 明 *Ming,* 清 *Qing,* 民國 *Republican period*), we extract all occurrences of a target lexeme w from the diachronic corpus. We process only texts with punctuation marks to ensure reliable sentence boundaries, using individual sentences as the contextual unit, aiming for contextual accuracy. Each occurrence is represented as a contextualized embedding $\mathbf{e}_{w,j}^{(i)} \in \mathbb{R}^{768}$ generated by **GuwenBERT**, a pre-trained language model specifically designed for Classical Chinese texts. GuwenBERT's training on large-scale historical corpora makes it particularly suitable for capturing the semantic nuances of pre-modern Chinese across different periods.

Two-Stage Clustering and Visualization. Given the set of embeddings $\mathcal{E}_i = \{\mathbf{e}_{w,1}^{(i)}, \ldots, \mathbf{e}_{w,n_i}^{(i)}\}$ for dynasty T_i, we employ a two-stage approach:

1. **K-means Clustering:** We first perform K-means clustering to partition $\mathcal{E}_i$ into k_i clusters $\mathcal{C}_i = \{c_{i,1}, \ldots, c_{i,k_i}\}$ in the full 768-dimensional embedding space. Each cluster $c_{i,j}$ represents a distinct computational sense of lexeme w in dynasty T_i.
2. **PCA Visualization:** After clustering, we apply Principal Component Analysis (PCA) to reduce the dimensionality to 2D for visualization and interpretation purposes. Importantly, the clustering is performed before dimensionality reduction to preserve the semantic relationships captured in the high-dimensional space.

The optimal number of clusters k_i is determined by the **Elbow method**[1], which we found through empirical validation to provide the most linguistically interpretable results compared to alternative metrics (Silhouette Coefficient, Gap Statistic, and BIC).

Cluster Characterization. Each cluster $c_{i,j}$ is characterized by:

- **Centroid vector:** The mean of all member embeddings, which provides interpretability and serves as an anchor to locate representative contexts.
- **Cluster size:** $|c_{i,j}|$, indicating sense frequency.
- **Variance:** Intra-cluster dispersion as a measure of sense coherence.
- **Representative contexts:** The sentences closest to the cluster centroid.

3.2. Cross-Dynasty Sense Alignment

Given two consecutive dynasties T_i and T_{i+1} with sense sets $\mathcal{C}_i$ and $\mathcal{C}_{i+1}$, we establish sense correspondences through average pairwise distance computation.

[1] The Elbow method identifies the point where adding additional clusters yields diminishing returns in variance reduction, corresponding well with human judgments of sense granularity.

$N \times M$ **Average Pairwise Distance.** For clusters $c_{i,a}$ from dynasty T_i (containing N embeddings) and $c_{i+1,b}$ from dynasty T_{i+1} (containing M embeddings), we compute:

$$d(c_{i,a}, c_{i+1,b}) = \frac{1}{N \cdot M} \sum_{p=1}^{N} \sum_{q=1}^{M} d_{\text{cos}}\left(\mathbf{e}_p^{(i,a)}, \mathbf{e}_q^{(i+1,b)}\right), \tag{1}$$

where

$$d_{\text{cos}}(\mathbf{u}, \mathbf{v}) = 1 - \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}. \tag{2}$$

This $N \times M$ approach captures the full distributional overlap between cluster populations, providing more robust alignment than centroid-only methods. The computational cost is justified by the improved accuracy in identifying subtle semantic shifts.

Unidirectional Alignment. For each cluster $c_{i,a}$ in dynasty T_i, we identify its successor in T_{i+1} as:

$$c_{i+1,b^*} = \arg \min_{c_{i+1,b} \in \mathcal{C}_{i+1}} d(c_{i,a}, c_{i+1,b}). \tag{3}$$

This creates directed edges in a diachronic sense graph, where convergence (multiple sources mapping to one target) indicates sense merger, and divergence suggests sense differentiation.

4. Pilot Study and Expected Results

To evaluate the feasibility of the proposed D-CWN pipeline, we conducted a focused pilot analysis on the lexical item **手** (hand), selected for its high token frequency, stable orthography across periods, and rich semantic diversity from physical to metaphorical uses.

4.1. Pilot Methodology

1. **Data Selection:** We extracted all occurrences of **手** (hand) from five randomly selected texts in the **Chinese Text Project (ctext)** database for each of the eight major historical periods: 戰國 (Warring States), 秦漢 (Qin–Han), 魏晉 (Wei–Jin), 隋唐 (Sui–Tang), 宋元 (Song–Yuan), 明 (Ming), 清 (Qing), 民國 (Republican period). This sampling strategy covers all major dynastic periods and likely reflects a range of genres in each dynastic period.

2. **Contextual Embedding:** Each occurrence of **手 (hand)** was tokenized using the **GuWenBERT** model, a transformer pretrained on large-scale pre-modern Chinese corpora. The model generates a contextualized vector representation for every instance, capturing the semantic features specific to its historical context.

3. **Clustering:** Within each dynasty, the contextual embeddings of **手 (hand)** were grouped using **k-means** clustering to create interpretable sense partitions. Each contextual embedding cluster approximates a distinct contextualized usage pattern or lexical sense.

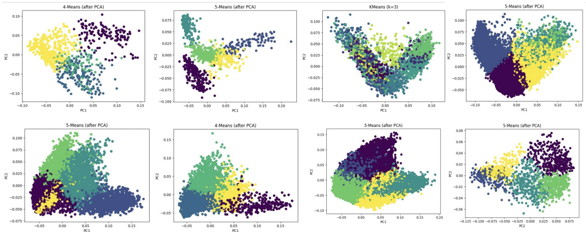

Figure 1: Clusters from 8 Consecutive Dynasties

4.2. Clustering Results

Figure 1 visualizes k-means clustering results for 手 (hand) across eight historical periods, with each panel corresponding to one period (from left to right, top to bottom: Warring States, Qin–Han, Wei–Jin, Sui–Tang, Song–Yuan, Ming, Qing, and Republican). Each point represents a single occurrence of 手 (hand) encoded as a GuwenBERT contextual embedding; clustering is performed in the original 768-dimensional space, and the points are then projected to two dimensions with PCA solely for visualization. Colors indicate cluster assignments, with panels illustrating different choices of k (e.g., 3, 4, or 5); titles such as "4-Means/5-Means (after PCA)"denote the selected cluster count and that the plotted coordinates come from the PCA projection.

4.3. Analysis & Discussion

Semantic stability is reflected in the clustering patterns themselves. When clusters are well separated, the senses of 手 (hand) appear more stable with clear category boundaries. When clusters substantially overlap, boundaries are fuzzier and meanings blend across contexts. With a higher k, a configuration that produces several dispersed groupings—especially if one cluster stands apart —points to broader semantic spread with a specialized submode. Conversely, tighter, more compact groupings indicate weaker sense differentiation and concentration around a smaller set of usages.

Across periods, these patterns trace diffusion versus convergence: When a period's clusters are more dispersed than in the subsequent period, this indicates semantic aggregation (consolidation of uses); conversely, when a period's clusters are less dispersed than in the subsequent period, this indicates semantic diffusion (expansion and diversification of uses).

5. Conclusion

In this work, we present the preparatory steps toward automating the construction of a Diachronic Chinese WordNet(D-CWN)., directly addressing the absence of a large-scale, temporally-aware lexical resource for historical Chinese. Our proposed three-stage pipeline successfully integrates contextualized embeddings from a specialized Classical Chinese language model, K-means clustering for within-dynasty sense discovery, and is expected to perform a computationally efficient centroid-based alignment for tracking sense evolution across historical periods. The pilot study on the character 手 validated our approach, and could further demonstrates that the use of cosine distance between sense centroids effectively identifies and quantifies patterns of semantic continuity, shift, and convergence in the future, laying a solid foundation for the full-scale construction of the D-CWN.

The contributions of this research are twofold: first, we have designed a scalable and reproducible pipeline that bridges the gap between statistical semantic change detection and structured lexicography; second, we have established a novel data-driven framework for classifying evolutionary sense relationships. While acknowledging the linguistic challenges of graphic variation and corpus-specific biases, our methodology is designed to be robust and adaptable. Future work will conduct computationally efficient alignment and focus on expanding our pilot study to a comprehensive lexicon of thousands of Chinese characters and incorporating human expert validation to produce a rich, queryable resource. Ultimately, the D-CWN promises to be an invaluable tool for researchers in Chinese NLP, digital humanities, and historical linguistics, enabling new lines of inquiry into the deep semantic history of the Chinese language.

6. Bibliographical References

Asaf Amrami and Yoav Goldberg. 2019. Towards better substitution-based word sense induction.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. *DIACR-Ita @ EVALITA2020: Overview of the EVALITA2020 Diachronic Lexical Semantics (DIACR-Ita) Task*, pages 411–419.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.

Jiahuan Cao, Yongxin Shi, Dezhi Peng, Yang Liu, and Lianwen Jin. 2024. C^3bench: A comprehensive classical chinese understanding benchmark for large language models.

Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training temporal word embeddings with a compass.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Chu-Ren Huang, Ru-Yng Chang, and Hsiang-Pin Lee. 2004. Sinica BOW (bilingual ontological Wordnet): Integration of bilingual WordNet and SUMO. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. 中文词汇网络: 跨语言知识处理基础架构的设计理念与实践 chinese wordnet: design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese information processing 1003-0077*, 24:14–23.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lung-Hao Lee, Shu-Kai Hsieh, and Chu-Ren Huang. 2009. CWN-LMF: Chinese WordNet in the Lexical Markup Framework. In *Proceedings* *of the 7th Workshop on Asian Language Resources (ALR7)*, pages 123–130, Suntec, Singapore. Association for Computational Linguistics.

Bin Li, Bolin Chang, Zhixing Xu, Minxuan Feng, Chao Xu, Weiguang Qu, Si Shen, and Dongbo Wang. 2024. Overview of EvaHan2024: The first international evaluation on Ancient Chinese sentence segmentation and punctuation. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 229–236, Torino, Italia. ELRA and ICCL.

Bin Li, Yiguo Yuan, Jingya Lu, Minxuan Feng, Chao Xu, Weiguang Qu, and Dongbo Wang. 2022. The first international Ancient Chinese word segmentation and POS tagging bakeoff: Overview of the EvaHan 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 135–140, Marseille, France. European Language Resources Association.

Mingchen Li, Zili Zhou, and Yanna Wang. 2020. Multi-fusion chinese wordnet (mcw) : Compound of machine learning and manual correction.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1003–1011, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2019. Survey of computational approaches to lexical semantic change.

Shan Wang and Francis Bond. 2013. Building the Chinese open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18, Nagoya, Japan. Asian Federation of Natural Language Processing.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

7. Language Resource References

Ethan-yt. 2020. Guwenbert. `https://github.com/Ethan-yt/guwenbert`. GitHub repository. Please cite the specific tag/commit if possible.

Donarld Sturgeon. 2011. Chinese text project: a dynamic digital library of premodern chinese. `https://ctext.org/`.

Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2021. Anchibert: A pre-trained model for ancient chineselanguage understanding and generation.