# Validating the Tails: Domain-Dependent Semantic-Axis Retrieval in Historical News

Yufei Zhu
The George Washington University
Washington, DC, USA
louisez@gwu.edu

## Abstract

Semantic axes provide an interpretable mechanism for ranking large text collections along conceptual dimensions, but the meanings assigned to their extreme scores are often inferred from the axis labels themselves. We examine whether such interpretations hold empirically when semantic axes are used for exploratory retrieval in historical news archives.

Using *New York Times* coverage from 1971–1989, we construct an escalation–détente axis and evaluate its extreme-score regions in two ways. First, blind human annotation of Soviet-related sentences shows that the high-scoring tail substantially enriches escalation-related discourse: 52% of high-tail sentences are labeled escalation, compared with 20% in a middle-score control (OR = 4.33, $p = .0016$). The corresponding low-scoring tail, however, does not significantly enrich détente, revealing asymmetric retrieval behavior.

Second, we freeze the embedding model and axis and apply them to Canada-related coverage. GPT-assisted topical coding of 400 tail sentences reveals sharply different semantic organization: the high tail is dominated by industry and infrastructure, whereas the low tail is dominated by politics and culture ($\chi^2(6) = 228.0$, permutation $p < .0001$, Cramér's $V = .755$).

These results suggest that semantic axes can support tail-sensitive exploratory retrieval, but their extreme regions should be empirically validated rather than interpreted directly from pole labels.

## 1 Introduction

Historical news archives contain collections too large for exhaustive reading. Keyword search provides effective access when researchers know the entities or events of interest, but exploratory research often begins with less precise questions about rhetoric, framing, or conceptual organization [4]. Embedding-based representations offer an alternative means of navigating such collections through continuous semantic structure [8].

Semantic axes are particularly attractive for this purpose because they represent semantic properties as directions between opposing concepts in embedding space [1, 2]. Documents or passages can then be ranked according to their projection onto such a direction [1]. A methodological problem follows, however: an axis named after two concepts does not guarantee that its extreme retrieval regions instantiate those concepts in context. Distributional representations are sensitive to the corpora and contexts from which they are learned [1, 3]. This raises a retrieval question that is less often tested directly: whether geometric opposition between axis poles corresponds to transferable rhetorical categories in context.

We investigate this problem using an escalation–détente axis over historical *New York Times* coverage. Rather than asking whether semantic-axis scores constitute a general classifier of Cold War rhetoric, we ask a narrower retrieval question: *what do the extreme regions of a semantic axis actually retrieve?*

We evaluate this question through two complementary experiments. Within Soviet-related coverage, blind human annotation tests whether high- and low-scoring regions enrich the rhetorical categories suggested by the axis. We then freeze the representation and transfer it to Canada-related coverage to examine how the semantic composition of the same score extremes changes outside the target domain.

Our results show asymmetric within-domain validity and strong domain dependence. The escalation tail significantly enriches escalation discourse, whereas the détente tail does not comparably enrich détente. Under transfer to Canadian coverage, the same axis instead separates industrial and infrastructural content from political and cultural content. We therefore argue that semantic axes are best treated as *tail-sensitive exploratory retrieval probes whose empirical interpretation requires domain-specific validation.*

## 2 Method

### 2.1 Corpus and Retrieval Units

We collect *New York Times* Archive API records from 1971–1989 [7]. Because metadata availability varies across years, we use the `abstract` field as the consistent textual representation. The resulting collection contains 1,243,711 records with usable abstracts.

We split abstracts into sentence-like units and retain sentences of at least 30 characters. For the target corpus, we retrieve sentences containing a predefined Soviet-related lexicon including references such as *USSR*, *Soviet*, *Soviet Union*, *Moscow*, and *Kremlin*. This procedure yields 46,282 Soviet-related sentences.

For the domain-transfer control, we apply the same sentence extraction procedure to Canada-related references, including *Canada*, *Canadian*, *Canadians*, and *Ottawa*.

### 2.2 Semantic Axis Construction

Following SemAxis [1], we construct an escalation–détente direction in a Word2Vec embedding space. The model is trained on the Soviet-related corpus using Word2Vec [5], with 200-dimensional vectors, a context window of 5, minimum frequency of 20, skip-gram training, and five epochs.

We initialize the two poles with *escalation* and *detente*. Following the neighbor-augmentation procedure of SemAxis [1], each seed is expanded with its 20 nearest neighbors in the trained embedding space. The escalation pole contains terms such as *instability*, *provoke*, and *deterrence*, while the détente pole contains terms such as *accommodation*, *relaxation*, and *coexistence*. Let $P$ and $N$ denote the resulting 21-word escalation and détente sets, respectively. We

| Year | Records | Year | Records |
|---|---|---|---|
| 1971 | 68,883 | 1981 | 56,523 |
| 1972 | 71,408 | 1982 | 66,764 |
| 1973 | 61,920 | 1983 | 79,294 |
| 1974 | 51,693 | 1984 | 74,703 |
| 1975 | 75,699 | 1985 | 71,237 |
| 1976 | 50,434 | 1986 | 82,987 |
| 1977 | 59,635 | 1987 | 72,796 |
| 1978 | 48,518 | 1988 | 70,732 |
| 1979 | 53,103 | 1989 | 78,421 |
| 1980 | 48,961 | | |

**Table 1: NYT Archive API records with usable abstracts, 1971–1989. The resulting collection contains 1,243,711 records; Soviet-related sentence filtering retains 46,282 sentences.**

define the axis as the difference between their centroids:

$$a = \frac{1}{|P|} \sum_{w \in P} v_w - \frac{1}{|N|} \sum_{w \in N} v_w,$$

where $v_w$ denotes the embedding of word $w$. The resulting vector is normalized and frozen for all subsequent experiments.

Each sentence is represented by the mean of its in-vocabulary word embeddings. Its axis score is the cosine similarity between the sentence representation $e_s$ and the normalized axis:

$$\text{score}(s) = \cos(e_s, a).$$

Higher and lower values indicate stronger geometric alignment with the escalation and détente directions, respectively. Importantly, we treat these values as retrieval scores, not rhetorical labels.

The complete expanded pole vocabularies, retrieval lexicons, and experiment configurations are retained with the experiment artifacts.

## 2.3 Experiment 1: Within-Domain Tail Validation

We first test whether score extremes correspond to their intended rhetorical interpretations in Soviet-related coverage. We define HIGH as the top 5% of the empirical Soviet score distribution, LOW as the bottom 5%, and MIDDLE as the 45th–55th percentile interval. We randomly sample 50 sentences from each stratum ($N = 150$).

The sentences are annotated blindly by a human annotator as escalation (E), détente (D), or neither/neutral (N), without access to their score or stratum. We compare the conditional probabilities $P(E \mid \text{HIGH})$ and $P(E \mid \text{MIDDLE})$, and analogously $P(D \mid \text{LOW})$ and $P(D \mid \text{MIDDLE})$. Fisher's exact tests quantify enrichment relative to the middle-score baseline.

## 2.4 Experiment 2: Domain-Transfer Control

We next freeze both the Word2Vec model and semantic axis and apply them without retraining to Canada-related sentences. Canada sentences are represented using only tokens present in the frozen Soviet-trained Word2Vec vocabulary. All 12,705 Canada-related sentences contain at least one in-vocabulary token and are therefore scorable. This experiment asks whether the same geometric extremes retain the interpretation suggested by the original axis.

| Stratum | E | N | D |
|---|---|---|---|
| HIGH | 26 (.52) | 20 (.40) | 4 (.08) |
| MIDDLE | 10 (.20) | 35 (.70) | 5 (.10) |
| LOW | 7 (.14) | 36 (.72) | 7 (.14) |

**Table 2: Blind human validation of Soviet-related retrieval strata ($n = 50$ per stratum). Values are counts with conditional proportions in parentheses. Escalation is enriched in HIGH relative to MIDDLE (OR=4.33, Fisher's exact $p = .0016$); détente is not significantly enriched in LOW relative to MIDDLE (OR=1.47, $p = .760$).**

We define HIGH and LOW using the Canada score distribution itself, as its top and bottom 5%, respectively, and randomly sample 200 sentences from each tail ($N = 400$). Using a fixed annotation protocol, GPT-4o-mini [6] assigns each sentence exactly one topical label: Industry/Business/Resources (IND), Infrastructure/Transport/Energy (INF), Accident/Disaster (ACC), Politics/Diplomacy/Government (POL), Culture/Arts/Media (CUL), Sports (SPT), or Other/Unclear (OTH). The model receives sentence text only and does not observe score or stratum.

We compare $P(\text{category} \mid \text{HIGH})$ with $P(\text{category} \mid \text{LOW})$. Because two expected cells contain fewer than five observations, we supplement the chi-square statistic with a 10,000-iteration permutation test. Cramér's $V$ measures the strength of association between retrieval tail and topical composition.

# 3 Results

## 3.1 The Target-Domain Axis Is Asymmetric

Human annotation shows that high scores meaningfully enrich escalation discourse. Of HIGH sentences, 52% are labeled escalation, compared with 20% of MIDDLE sentences. This corresponds to an odds ratio of 4.33 (Fisher's exact $p = .0016$).

The opposite end of the axis does not exhibit comparable behavior. Only 14% of LOW sentences are labeled détente, compared with 10% in MIDDLE (OR = 1.47, $p = .760$). Neutral/other interpretations remain dominant in both the MIDDLE (70%) and LOW (72%) strata.

Thus, geometric opposition between the two axis poles does not imply symmetric rhetorical retrieval. The positive tail provides a useful escalation-oriented retrieval region, whereas the negative tail cannot be interpreted straightforwardly as a détente detector.

## 3.2 The Frozen Axis Changes Meaning Under Transfer

The Canada control produces a substantially different organization of the score extremes. In the HIGH tail, 58.0% of sentences concern industry/business/resources, followed by infrastructure/transport/energy (15.5%) and politics (12.0%). In contrast, the LOW tail is dominated by politics/diplomacy/government (49.5%) and culture/arts/media (38.5%); only 10.5% concern industry.

The overall topical distributions differ strongly between tails ($\chi^2(6) = 228.0$). Because two expected cells fall below five, we assess significance using a 10,000-iteration permutation test, which

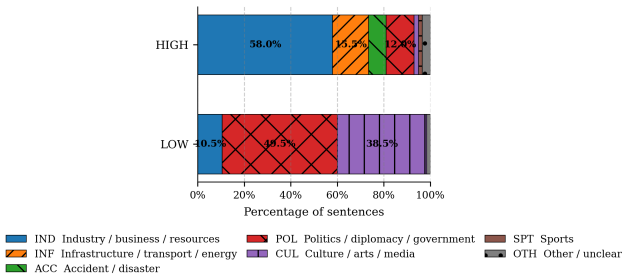



**Figure 1: Topical composition of the HIGH and LOW retrieval tails in Canada-related coverage ($N = 400$). The frozen Soviet-derived axis separates an industry-dominated HIGH tail from a LOW tail dominated by politics and culture, illustrating that the empirical interpretation of a semantic axis can shift under domain transfer.**

confirms the association ($p < .0001$); the effect is large (Cramér's $V = .755$).

These results operationalize domain dependence as a change in the empirical interpretation of score extremes under transfer. The numerical direction remains fixed, but the content organizing its tails changes with the retrieval domain.

## 4 Discussion and Conclusion

Our experiments highlight two cautions for semantic-axis retrieval in historical archives.

First, axis geometry is not necessarily semantically symmetric. Although the escalation end of our axis significantly enriches escalation-related discourse, the détente end does not comparably enrich détente. Naming an axis after two opposing poles therefore does not establish that both extremes function as reliable rhetorical categories.

Second, axis interpretation is domain-dependent. When the same frozen representation is applied to Canadian coverage, the extremes organize substantially different topical distinctions: industrial and infrastructural material concentrates toward the high end, while political and cultural material concentrates toward the low end. A semantic direction that supports useful retrieval in one domain should therefore not be assumed to preserve the same interpretation elsewhere.

These findings suggest a practical role for semantic axes in digital-library interfaces. Rather than treating projection scores as classifications, systems can expose them as exploratory rankings whose tails are inspected and validated against the collection being studied. Such probes remain lightweight and interpretable while avoiding stronger claims than their empirical behavior supports.

Our study is limited to one newspaper archive, one semantic axis, and a single human annotator for the target-domain evaluation. The Canada control additionally uses model-assisted rather than human topical annotation. Future work should evaluate inter-annotator reliability, additional collections and conceptual axes, and interfaces that expose domain-specific tail behavior directly to archival researchers.

Overall, semantic axes appear useful not because their pole labels universally determine meaning, but because their extreme regions can reveal coherent, collection-specific structures. For exploratory archival retrieval, the important question is therefore not simply how an axis is named, but what it actually retrieves in a particular collection.

## References

[1] Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*. Association for Computational Linguistics, 2450–2461. doi:10.18653/v1/P18-1228

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[3] 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)*. 1489–1501.

[4] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[6] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[7] The New York Times Company. 2026. The New York Times Developer API. https://developer.nytimes.com/. Accessed: 2026-09-15.

[8] Janus Wawrzinek, Said Ahmad Ratib Hussaini, Oliver Wiehr, José María González Pinto, and Wolf-Tilo Balke. 2020. Explainable word-embeddings for medical digital libraries-a context-aware approach. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. 299–308.