Automatic Construction of a Patent Term Thesaurus with Fine-Tuned ChatGPT

Hidetsugu Nanba[1], Kohei Iwakuma[1] and Satoshi Fukuda[1]

[1]Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551, Japan

Abstract

Technical terms in patent documents are expressed with highly variable, domain-specific language, hindering cross-lingual prior-art search and knowledge discovery. Existing automatic thesaurus construction pipelines either rely on handcrafted patterns, which suffer from low recall, or on graph-augmented representation learning, which is accurate but complex and largely monolingual. We present a lightweight three-stage framework that: **(1)** filters candidate term pairs with off-the-shelf embeddings, **(2)** assigns fine-grained semantic relations via a ChatGPT-4o model fine-tuned on 36k English patent pairs, and **(3)** enforces cross-lingual consistency through fixed-expression hypernym seeds automatically aligned between Japanese and English. The final output is written directly into an incrementally updateable multilingual thesaurus graph. On the Google Patent Phrase Similarity Dataset, our fine-tuned LLM attains 0.762 Pearson / 0.738 Spearman, outperforming strong baselines (SBERT, Patent-BERT) and the recent graph-based *RA-Sim* model by up to 0.14 correlation points.

Keywords

Large Language Models, Term Relation Extraction, Thesaurus Construction

1. Introduction

Patent documents constitute a rich, internationally distributed repository of technical knowledge; yet the diversity of languages, orthographies, and complex compound terms poses a formidable barrier to cross-lingual search and comparison. National patent offices and companies therefore spend substantial resources to maintain thesauri and classification schemes. Previous attempts at automating this process have focused on pattern-based extraction (e.g. Hearst patterns[1]) or supervised models built on Word2Vec[2]/BERT-style embeddings[3], but their single-language assumptions and limited domain adaptation have left the multilingual coverage of specialised terms and the capture of fine-grained semantic relations (hypernymy, meronymy, etc.) still inadequate.

This paper leverages the broad cross-lingual knowledge encoded in OpenAI's ChatGPT-4o to introduce a simple framework that, given any pair of patent terms, directly predicts their semantic relation. High accuracy in English is ensured by fine-tuning the model on the Google Patent Phrase Similarity Dataset[4], while extension to other languages is achieved without additional corpora, simply by exploiting the model's latent multilingual representations. For empirical verification, we extract hypernym–hyponym candidates from Japanese and English patents using fixed expressions such as "A nado no B" (Japanese) and "B such as A" (English), align them through automatic translation, and use the resulting high-confidence bilingual pairs to test whether the model assigns consistent relations across languages.

The proposed study makes three contributions.

- It demonstrates that a large language model can be transferred to multilingual specialised-term relation prediction with only minimal fine-tuning, drastically reducing the cost of building and maintaining separate models for every language.
- It presents a workflow that organises the relations predicted by the LLM into a graph whose nodes are terms and edges are relations, and then expands this graph recursively to build an automatically updatable multilingual patent thesaurus.
- It introduces an evaluation procedure that combines pattern-based hypernym candidates extracted independently from Japanese and English patents with their translation alignments, enabling the community to verify whether the LLM's predictions remain consistent across languages.

2. Related Work

Automatic prediction of semantic relations—synonymy, hypernymy, meronymy, and the like—between technical terms has long underpinned knowledge acquisition and high-recall retrieval. Historical approaches fall into three broad families: *symbolic pattern rules*, *distributional or embedding methods*, and, most recently, *large language models (LLMs)*. Below we survey their evolution in chronological order, emphasising patent-specific work and highlighting how our study differs.

Early research relied on explicit lexico-syntactic patterns. Hearst's seminal paper introduced templates such as "*X is a kind of Y*" to harvest thousands of hypernym–hyponym pairs at negligible cost [1]. The same idea was later applied to Japanese patent corpora: Nanba et al. mined the pattern "A nado no B" (B such as A), aligned the resulting pairs with English equivalents, and built a bilingual thesaurus with 78 % F_1 [5]. Building on this, their subsequent study translated scholarly terms into patent terminology by combining citation analysis with an automatically constructed thesaurus, significantly broadening the candidate space [6]. Their scope, however, is restricted to *hypernym–hyponym* relations only, whereas the present study predicts a full spectrum of relations—including synonymy, meronymy, and graded similarity—across languages. Symbolic methods moreover demand handcrafted patterns for every language and domain; even in English, Roller et al. revisited Hearst rules with modern corpora to boost accuracy, yet still faced recall limits when wording drifted from canonical templates [7]. Patents exacerbate this problem:

6th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech) 2025

nanba@kc.chuo-u.ac.jp (H. Nanba);
fukuda.satoshi.3238@kc.chuo-u.ac.jp (S. Fukuda)
http://nlp.ds.chuo-u.ac.jp/index_en.html (H. Nanba)
0000-0002-0877-7063 (H. Nanba)

identical concepts are phrased idiosyncratically ("soccer ball'' vs. "spherical recreational device''), so surface patterns alone capture only a fraction of true relations. A broader survey of how rule-based and other NLP techniques transfer—or fail to transfer—between patent sub-genres is given by Andersson et al. [8]. Complementary work by Judea et al. shows that figure references themselves can be harvested as symbolic cues, yielding fully unsupervised, high-quality training data for patent terminology extraction [9].

Distributional approaches learn continuous vectors from large corpora. Word2Vec[2] and GloVe[10] established that words with similar contexts occupy nearby positions in an embedding space. Jana et al. projected a distributional thesaurus into such a space and achieved strong co-hyponym detection by clustering context-similar terms [11]. However, plain similarity cannot distinguish type (synonymy versus hypernymy). Subsequent work trained classifiers or added constraints; Liu et al. prompted BERT with masked templates ("*X is a type of* __'') to recover hypernyms more robustly [12]. Contextual models improved further with Transformer pre-training: BERT [3] and its Siamese variant Sentence-BERT (SBERT) [13] achieved state-of-the-art semantic similarity. Yet domain adaptation proved essential—Patent-BERT, trained on claim corpora, vastly outperformed general BERT on patent relation benchmarks.

The advent of LLMs enabled direct reasoning over relations. Models such as ChatGPT-4 store vast world knowledge and can generate definitions, synonyms, or hypernyms with minimal prompting. Recent reports show ChatGPT-4 successfully deriving taxonomic links for multilingual cultural terms, indicating latent cross-lingual competence unavailable to earlier systems. In the patent realm, Peng and Yang combined a contextual encoder with a citation-derived phrase graph; their self-supervised method captured global evidence beyond local context and raised similarity correlation by seven points [14]. Such hybrids improve accuracy but demand heavy pipelines (citation crawling, graph learning) and remain monolingual.

Cross-domain evaluation has been invigorated by resources tailored to patents. The Google Patent Phrase Similarity Dataset supplies 50 k phrase pairs with graded similarity and relation labels [4]; Kaggle competitions around it confirmed SBERT-style models as strongest baselines and revealed the benefit of patent-specific pre-training. Yet most entries handled English only and did not automate thesaurus induction.

Our study departs from prior art in three ways. First, we retain a lightweight embedding filter but rely on a minimally fine-tuned ChatGPT-4o to infer relations, avoiding bespoke citation graphs or rule sets. Second, we enforce cross-lingual consistency via pattern-harvested bilingual seed pairs, allowing the same model to populate a thesaurus in Japanese and English without extra translation resources. Third, the LLM's output is written directly into an incrementally expandable graph, turning relation inference into immediate thesaurus construction rather than a separate post-processing step. In doing so, we address the lingering gaps of multilingual coverage, domain knowledge acquisition, and pipeline complexity that earlier approaches left open.

3. Proposed Method

Our framework builds a multilingual patent thesaurus through two *alternative* relation–inference strategies plus a multilingual verification step: **(i) an embedding-based *similarity* inference**, **(ii) an LLM-based *explicit-label* inference**, and **(iii) pattern-driven multilingual enrichment**. Stages (i) and (ii) pursue the *same objective*—predicting the semantic relation of a term pair—but differ in the signal they exploit: dense vectors vs. generative reasoning. Stage (iii) then enforces cross-lingual consistency and incrementally expands the thesaurus graph.

3.1. Embedding-Based Similarity Inference

Given a term pair (t_i, t_j), we obtain vectors $\mathbf{e}_i, \mathbf{e}_j \in \mathbb{R}^d$ from either OpenAI Embeddings ($d = 1536$) or multilingual-e5-large[1] ($d = 1024$). Their cosine similarity,

$$\mathrm{sim}(t_i, t_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \, \|\mathbf{e}_j\|},$$

serves as a *proxy score* for semantic relatedness. Pairs whose score exceeds a threshold τ (0.35 for OpenAI, 0.30 for e5) are tentatively regarded as *related* (synonym or taxonomic) and forwarded to the multilingual verification in Stage (iii). This embedding view offers a fast, language-agnostic approximation that requires no fine-tuning.

3.2. LLM-Based Explicit Relation Inference

Alternatively, the same pair can be passed to ChatGPT-4o mini, fine-tuned on the *Google Patent Phrase Similarity Dataset*. The prompt asks:

> Based on ' reading machine', what is the relationship of ' photocopier'? Please choose the most appropriate one from the following:
> 1: 'Not related.'
> 2: 'Other high level domain match.'
> 3: 'Holonym (a whole of).'
> 4: 'Meronym (a part of).'
> 5: 'Antonym.'
> 6: 'Structural match.'
> 7: 'Hypernym (narrow-broad match).'
> 8: 'Hyponym (broad-narrow match).'
> 9: 'Highly related.'
> 10: 'Very highly related.'

The model chooses a single label from Table 1; we map it to a numerical score $\{1.00, 0.75, 0.50, 0.25, 0.00\}$. Compared with Stage (i), the LLM returns an explicit relation type (e.g., Hyponym, Meronym) rather than a scalar similarity.

3.3. Pattern-Driven Multilingual Enrichment

1. **Seed extraction**:
   - *Japanese*: phrases matching "A nado no B"
   - *English*: phrases matching "B such as A"

   These patterns produce provisional hyponym (A) / hypernym (B) pairs.
2. **Translation alignment**: English pairs are machine-translated to Japanese using ChatGPT and intersected with the Japanese set;*high-confidence bilingual pairs.*

[1] https://huggingface.co/intfloat/multilingual-e5-large

Table 1
Relation labels and their scores.[4]

Relation label	Score
Very Highly related	1.00
Highly related	0.75
Hyponym / Hypernym	0.50
Structural match	0.50
Meronym / Holonym	0.25
Antonym	0.25
Other domain match	0.25
Not related	0.00

3. **Cross-lingual verification**: Each pair is checked by either Stage (i) or (ii); only pairs whose Japanese and English predictions agree are accepted.
4. **Thesaurus graph update**: Accepted pairs become edges (relation type) between term nodes. The graph updates automatically as new pairs arrive.

By offering two complementary inference routes—fast embedding similarity or explicit LLM labelling—and a verification layer that fuses them across languages, our method achieves multilingual coverage with minimal fine-tuning while avoiding complex citation graphs or handcrafted rules. Experimental details follow in Section 4.

4. Experiments

4.1. Experimental Setup

Datasets For the English task we adopt the *Google Patent Phrase Similarity Dataset*, using 36,473 pairs for training and 9,232 for validation and testing.

Alternatives

- **Embedding models:** Word2Vec, GloVe, BERT, SBERT, Patent-BERT (baselines reported by [4]), OpenAI Embeddings (text-embedding-3-large), and multilingual-e5-large.
- **Graph + encoder:** the phrase-graph embeddings released with *RA-Sim* (a baseline reported by [14]).
- **LLMs:** ChatGPT-4o and ChatGPT-4o mini in their pretrained form, plus fine-tuned versions on the English training split.

Metrics For English we report Pearson and Spearman correlation between predicted similarity scores and gold scores.

4.2. Results

The results are shown in Table 2. The fine-tuned ChatGPT-4o attains the strongest correlation (Pearson 0.762), outperforming the graph-augmented *RA-Sim* by 0.14 Pearson / 0.09 Spearman.

4.3. Discussion

To verify the effectiveness of fine-tuning, we compared similarity scores before and after adaptation. Table 3 shows that scores improved for **42%** of pairs with ChatGPT-4o and **52%** with ChatGPT-4o mini, while only 10 % deteriorated. The overall distribution shifted toward values closer to the gold standard, indicating that fine-tuning successfully supplements the model's domain knowledge and yields more accurate similarity estimates.

Table 2
Patent phrase similarity inference performance.

Type	Model	Pearson	Spearman
Embedding	Word2Vec[4]	0.437	0.483
	GloVe[4]	0.429	0.444
	BERT[4]	0.418	0.409
	SBERT (all-mpnet)[4]	0.598	0.535
	Patent-BERT[4]	0.528	0.535
	OpenAI Embeddings	0.581	0.564
	multilingual-e5-large	0.574	0.546
Graph +encoder	RA-Sim[14]	0.622	0.652
LLM	ChatGPT-4o	0.505	0.514
	ChatGPT-4o mini	0.371	0.403
	ChatGPT-4o (fine-tuned)	**0.762**	**0.738**
	ChatGPT-4o mini (fine-tuned)	0.742	0.718

Table 3
Change in similarity scores after fine-tuning.

	Improve	**Same**	**Impair**
ChatGPT-4o	3,899 (0.422)	4,335 (0.470)	998 (0.108)
ChatGPT-4o mini	4,806 (0.521)	3,520 (0.381)	906 (0.098)

Because the LLM classifies each pair into ten semantic relations, we can compute precision and recall for every class. Tables 4 and 5 list the fine-tuned ChatGPT-4o and ChatGPT-4o mini results, respectively. Both models excel at *Not related*, *Antonym*, and high-similarity classes, while *Holonym*, *Meronym*, and *Structural match* remain challenging—mainly due to their scarcity in the training data. Therefore, we constructed a multilingual thesaurus while improving these problems using the method proposed in Section 3.3.

Table 4
Evaluation results for each relation label by ChatGPT-4o (fine-tuned).

Relation	P	R	F_1
Very highly related	0.94	0.73	0.82
Highly related	0.50	0.58	0.54
Hypernym	0.42	0.44	0.43
Holonym	0.27	0.24	0.25
Structural	0.05	0.02	0.03
Meronym	0.18	0.20	0.19
Hyponym	0.41	0.40	0.41
Antonym	0.71	0.59	0.64
Other domain	0.33	0.30	0.3
Not related	0.74	0.79	0.76

5. Automatic Construction of a Multilingual Thesaurus Using Cross-Lingual Verification

We automatically construct a multilingual thesaurus from the full text of Japanese and US patents published between 1993 and 2023. Our main objective is to extract hypernym-hyponym relationships, but we also extract other relationships in the process. The procedure is described below.

1. Using the expressions "A nado no B" (Japanese) and "B such as A" (English), we extracted 613,251

Table 5
Evaluation results for each relation label by ChatGPT-4o mini (fine-tuned).

Relation	P	R	F_1
Very highly related	0.84	0.72	0.78
Highly related	0.47	0.49	0.48
Hypernym	0.41	0.38	0.39
Hyponym	0.42	0.43	0.43
Structural	0.03	0.03	0.03
Meronym	0.16	0.15	0.15
Holonym	0.25	0.29	0.27
Antonym	0.61	0.62	0.61
Other domain	0.31	0.31	0.31
Not related	0.74	0.76	0.75

Japanese and 518,166 English candidate pairs and kept 42,784 bilingual pairs after translation alignment using ChatGPT.

2. ChatGPT-4o mini (fine-tuned) predicted relations for both languages; only pairs with matching labels were retained (21,673 pairs).

In Step 2, we decided to use ChatGPT-4o mini (fine-tuned), which is comparable to ChatGPT-4o, which had the highest value in Table 2, because processing large amounts of data is extremely costly.

Tables 6 and 7 show the distribution of labels obtained by classifying the top and bottom candidates in English and Japanese from Step 1 using ChatGPT-4o mini (fine-tuned). Additionally, Table 8 shows the distribution of labels for the results where English and Japanese agree.

Table 6
Relation label distribution by ChatGPT-4o mini (English).

Relation	Count	Share
Very highly related	311	0.007
Highly related	1,626	0.038
Hypernym	3,419	0.080
Hyponym	26,075	0.610
Structural	258	0.006
Meronym	2,815	0.066
Holonym	3,345	0.078
Antonym	68	0.002
Other domain	3,438	0.080
Not related	1,429	0.033

Table 7
Relation label distribution by ChatGPT-4o mini (Japanese).

Relation	Count	Share
Very highly related	482	0.011
Highly related	2,642	0.062
Hypernym	4,309	0.101
Hyponym	24,138	0.564
Structural	394	0.009
Meronym	2,433	0.057
Holonym	3,448	0.081
Antonym	82	0.002
Other domain	3,571	0.084
Not related	1,285	0.030

Figure 1 shows a part of the multilingual thesaurus created using the proposed method. When 200 items were randomly selected and evaluated manually, 194 (97%) were correct. Among the incorrectly extracted entries, there were five cases that were not entirely incorrect but depended on context, such as “materials (anchor) - metals (target) - Hyponym (relation)” and “information (anchor) - time (target) - Hyponym (relation),” and one case that was completely incorrect, such as “materials (anchor) - combinations (target) - Homonym (relation).”

Table 8
Relation label distribution where Japanese and English predictions agree.

Relation	Count	Share
Very highly related	292	0.013
Highly related	468	0.022
Hypernym	556	0.026
Hyponym	18,355	0.847
Structural	3	0.000
Meronym	425	0.020
Holonym	622	0.029
Antonym	7	0.000
Other domain	724	0.033
Not related	221	0.010

Anchor (English)	Target (English)	Anchor (Japanese)	Target (Japanese)	Relation label
network	Internet	ネットワーク	インターネット	Hyponym
metal	aluminum	金属	アルミニウム	Hyponym
liquid	water	液体	水	Hyponym
vehicle	automobile	車両	自動車	Highly related

Figure 1: Examples of Japanese–English pairs whose predicted relation labels agreed.

6. Conclusion

We introduced a three–stage pipeline that combines a lightweight embedding filter, a minimally fine-tuned ChatGPT-4o, and pattern-driven cross-lingual verification to build a continuously expandable multilingual patent thesaurus. Experiments on the Google Patent Phrase Similarity Dataset demonstrated that the proposed LLM surpasses both embedding baselines and the recent graph-augmented *RA-Sim* model (Pearson 0.762 vs. 0.622). On 42,784 automatically aligned Japanese–English hypernym pairs, the pattern + LLM strategy achieved 97 % accuracy.

The framework requires no citation crawling, no external knowledge base, and no language-specific rules beyond a handful of fixed expressions, yet delivers state-of-the-art accuracy while remaining fully incremental. These traits make it attractive for industry settings where frequent thesaurus updates and multilingual coverage are essential.

References

[1] M. A. Hearst, Automatic acquisition of hyponyms from large text corpora, in: Proceedings of COLING ’92, 1992, pp. 539–545.

[2] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL 2019, 2019, pp. 4171–4186.

[4] G. Aslanyan, I. Wetherbee, Patents phrase to

phrase semantic matching dataset, arXiv preprint arXiv:2208.01171 (2022).

[5] H. Nanba, S. Mayumi, T. Takezawa, Automatic construction of a bilingual thesaurus using citation analysis, in: Proceedings of the PaIR'11 Workshop, 2011, pp. 1–8.

[6] H. Nanba, H. Kamaya, T. Takezawa, M. Okumura, A. Shinmori, H. Tanigawa, Automatic translation of scholarly terms into patent terms, in: Proceedings of the 2nd International Workshop on Patent Information Retrieval (PaIR '09), Association for Computing Machinery, Hong Kong, China, 2009, pp. 21–24. DOI forthcoming.

[7] S. Roller, D. Kiela, M. Nickel, Hearst patterns revisited: Automatic hypernym detection from large text corpora, in: Proceedings of ACL 2018, 2018, pp. 358–363.

[8] L. Andersson, A. Hanbury, A. Rauber, The portability of three types of text mining techniques into the patent text genre, in: Mihai Lupu, Katja Mayer, John Tait, Anthony J. Trippe (Eds.), Current Challenges in Patent Information Retrieval, volume 37 of *The Information Retrieval Series*, Springer, Berlin / Heidelberg, 2017, pp. 241–280. doi:10.1007/978-3-662-53817-3_9.

[9] A. Judea, H. Schütze, S. Brügmann, Unsupervised training set generation for automatic acquisition of technical terminology in patents, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin City University and the Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 290–300.

[10] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of EMNLP 2014, 2014, pp. 1532–1543.

[11] A. Jana, N. R. Varimalla, P. Goyal, Using distributional thesaurus embedding for co-hyponymy detection, in: Proceedings of LREC 2020, 2020, pp. 5766–5771.

[12] C. Liu, T. Cohn, L. Frermann, Seeking clozure: Robust hypernym extraction from bert with anchored prompts, in: Proceedings of *SEM 2023, 2023, pp. 193–206.

[13] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of EMNLP–IJCNLP 2019, 2019, pp. 3982–3992.

[14] Z. Peng, Y. Yang, Connecting the dots: Inferring patent phrase similarity with retrieved phrase graphs, in: Proceedings of NAACL 2024, 2024, pp. 1877–1890.