A graph-based analysis of semantic types and coercion in contextualized word embeddings

Long Chen **Deniz Ekin Yavaş**
Heinrich Heine University Düsseldorf
{chen.long,deniz.yavas}@hhu.de

Abstract

Semantic type mismatch between a noun and its context is central to coercion phenomena. This paper introduces a graph-based method to examine how lexical and contextual type information is reflected in word embeddings. We select nouns from ten semantic types, annotate corpus instances for type matching (matching vs. coercion vs. other mismatch vs. unrestricted), and construct graphs using BERT and sense-enhanced embeddings. Two metrics—Neighbor Type Probability (NTP) and Neighbor Type Entropy (NTE)—are proposed to analyze neighborhood type distributions. Results show that graphs constructed with sense-enhanced embeddings reflect semantic type information better, and matching and mismatch sentences can be distinguished through the proposed metrics.

1 Introduction

Word meaning is composed of various aspects, and semantic type is a fundamental part of it (e.g. Montague, 1970). It is related not only to the conceptual category of the noun but also its usage in the grammar. Ideally, semantic types can be defined by the distribution of words (Asher, 2011). For example, the noun 'pizza' in (1-a) belongs to the semantic type *food* and typically co-occurs with predicates such as 'delicious' and 'eat'.

(1) a. I am eating a delicious pizza.
b. I finished the pizza.

However, nouns are not always used in their prototypical, literal way. The context of an instance may require a semantic type different from that of the noun itself. Coercion (Pustejovsky, 1993) is a typical case of mismatch between the lexical type, i.e. the semantic type of an instance itself and the context type, i.e. the semantic type the context requires or suggests. In a coercion sentence like (1-b), the noun co-occurs with 'finish', which typically selects for an *activity* rather than a *food*.

In this paper, we investigate how the information about semantic types is reflected in the contextualized word embeddings of noun instances, both in typical cases of normal predication as in (1-a) and in non-canonical cases of coercion as in (1-b). For this purpose, we conduct a graph-based analysis of contextualized word embeddings using cosine similarity between the embeddings to connect instances similar to each other.

We construct a dataset with nouns from ten semantic types and annotate their corpus occurrences as different types of sentences. Four possible types of sentences are distinguished, with matching sentences and coercion sentences being the main focus of our study. A sentence is seen as a matching sentence if the lexical type matches the contextual type, and coercion is a typical case of mismatch between the lexical type and the contextual type.

We experiment with different language model embeddings: BERT embeddings (Devlin et al., 2019) and sense-enhanced embeddings (Yavas et al., 2025). The latter is a fine-tuned variant of BERT in order to incorporate WordNet supersense information (Fellbaum, 1998) into the embeddings. WordNet supersenses correspond well to different semantic types. In addition, we also create graphs using masked versions of these embeddings, where the target word is replaced with a [MASK] token. This allows the focus on the information about contextual types by removing lexical information. We propose two metrics (Neighbor Type Probability (NTP) and Neighbor Type Entropy (NTE)) to quantify the distribution and diversity of semantic types among each instance's neighbors in the graph.

By comparing the neighbor types across the instances within the same lexical type or the same sentence type, we discover that graphs constructed with sense-enhanced embedding reflect the seman-

tic types better than the ones with BERT embeddings; the lexical types are reliably reflected by the graphs constructed with sense-enhanced embeddings, while contextual types are also partially reflected by graphs constructed with masked embeddings. Instances in different types of sentences display a different pattern in terms of the types of their neighbors. In matching sentences like (1-a), the instances usually share the same lexical type as their neighbors, while in coercion sentences like (1-b) the instances exhibit a higher diversity of types among the neighbors. These findings indicate the effectiveness of our graph-based method.

2 Related work

2.1 Coercion

Coercion has been a key challenge to the theory of semantic types, as the lexical type of a noun in a coercion construction conflicts with the expected types from the context. A range of formal frameworks have been proposed to address this challenge, including Generative Lexicon (GL) (Pustejovsky, 1995), Type Compositional Logic (TCL) (Asher, 2015), Modern Type Theory (MTT) (Luo, 2012), frame semantics (Chen et al., 2022), among others.

These approaches can be broadly divided into two groups according to how they model the mechanisms underlying coercion. The first group, which includes GL and MTT, assumes a shift in the semantic type of the coerced instance itself. In GL, for example, a noun instance in coercion undergoes a type shift, as in (2), the type of the noun 'bottle' shifted from *artifact* to *activity*.

(2) I finished the bottle off in two gulps.

The second group, including TCL and frame semantics, assumes a different semantic type not on the coerced noun but on the predication relation, or more generally speaking, the context around the noun. Within frame semantics, for instance, in (2), the noun *bottle* retains its type *artifact*, and it is the predicate that accepts an *artifact* as its object and creates an eventive reading. Despite the richness of these theoretical proposals, computational work that directly supports or operationalizes them remains comparatively scarce. One notable exception is Asher et al. (2016), who applied distributional models to adjective–noun composition—a domain that includes coercive cases—and provided evidence in favor of TCL.

2.2 Pre-Trained Language Models and Semantic Types and Coercion

Few studies focus on whether semantic type knowledge is encoded in the embeddings of pre-trained language models. These studies train classifiers using the frozen embeddings of pre-trained language models and show high performance (Zhao et al., 2020; Yavas et al., 2023).

Several studies focus on coercion interpretation with pre-trained language models based on their word predictions and more specifically focusing on coercion to *event*. Rambelli et al. (2020) investigate covert event retrieval in cases of coercion in English, evaluating several pre-trained language models against human judgments. Their results show that the models fail to substantially outperform simpler distributional models.

Gietz and Beekhuizen (2022) show that coercion interpretation is not necessarily resolved to a single covert event. This is evidenced by low human annotator consensus on the underlying event for naturally-occurring coercion sentences in English. They test different computational models including BERT, co-occurrence counts, prototype vector, and example-based learning models. BERT's performance is tested using masked word prediction and it outperforms other computational models in both high and low consensus cases. Ye et al. (2022) evaluate coercion interpretation in English using masked word prediction, measuring whether the top-1 and top-3 predictions of BERT match the underlying covert event and they report poor performance.

Radaelli et al. (2025a) investigate covert event interpretation in coercion sentences in Norwegian using word prediction across 17 language models. They evaluate whether in models' predictions plausible events are ranked above less plausible ones. Results show that models fail to systematically and consistently rank plausible events higher, and only few outperform a simple corpus frequency baseline. Radaelli et al. (2025b), following Radaelli et al. (2025a), investigate how additional context affects coercion interpretation. While the models generally benefit from additional context, the performance varies depending on the model. Models that struggle in context-neutral sentences show greater improvements.

These studies investigate coercion focusing on word predictions of the pre-trained language models. To our knowledge, our study is the first to fo-

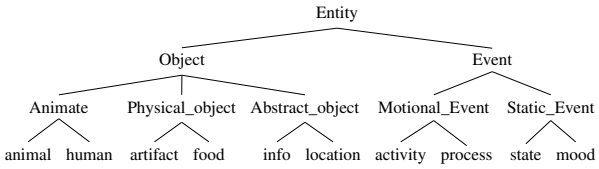

Figure 1: A presumed type hierarchy of the selected semantic types.

cus on this phenomenon on the representation level. Furthermore, these studies focus on one type of coercion. We examine a broader range of coercion types and investigate how semantic type relations are more generally captured by the contextualized word embeddings of BERT.

3 Data

We select a set $\mathcal{T}$ of ten semantic types for analysis: $\mathcal{T} = \{$*animal*, *artifact*, *activity*, *food*, *human*, *info*, *location*, *mood*, *process*, *state*$\}$. Some of them are ontologically related and can be grouped into higher-level concepts. For example, *human* and *animal* together form the superordinate category of *animate* entities, while *state* and *mood* can be subsumed under *static event*. A conceptual hierarchy of these ten semantic types is shown in Fig. 1. These types are selected because their prototypical instances are relatively easy to distinguish. Moreover, each type is presumably close (in the ontological sense) to some types but not others, which motivates a basic hierarchical structure. This hierarchy is broadly compatible to some existing taxonomies of the relevant types (e.g. WordNet (Fellbaum, 1998)) and could be tested with our method.

For each type, we select five to ten relatively frequent nouns. Some types have fewer nouns due to the limited number of high-frequency examples.[1] For each selected noun, we randomly extract 20 sentences containing the noun from BookCorpus (Zhu et al., 2015). A subset of sentences is manually removed for two reasons. First, some sentences exhibit a similar syntactic and semantic distribution to other existing sentences. Second, certain selected nouns are polysemous and the instance in the extracted sentence corresponds to another sense entry. After the filtering process, each noun is represented by 10 to 16 sentences. All the selected nouns can be found in the Table 7 and 8 in the appendix, and the full dataset will be published upon acceptance.

Each sentence is annotated with information about the semantic types related to the selected noun. Two notions of semantic types are distinguished: lexical type and contextual type (abbreviated as lt and ct). Lexical type refers to the semantic type of the noun itself, and contextual type refers to the semantic type that the surrounding context requires or suggests. In practice, the contextual type of an instance can be inferred by masking the noun in the sentence. Consider example (3-a), where the lexical type of the noun 'conference' is *activity*. When the noun is masked, as in (3-b), the noun that can felicitously fill the masked position can be 'meeting', 'match', 'argument', all of which are associated to *activity*. Therefore, the contextual type of 'conference' in (3-a) is also *activity*. Such sentences, where $lt = ct$, are annotated as MATCHING. In our dateset, 88% of the sentences (1410 sentences) are labeled as MATCHING, consistent with the general assumption that in a sentence, the contextual type matches the lexical type.

(3) a. The **conference** with James Stickley lasted for more than an hour. (lt: *activity*)
b. The [MASK] with James Stickley lasted for more than an hour.

In addition to the matching cases discussed above, we also encounter instances where $lt \neq ct$, i.e. the semantic type of the noun does not align with the type required or implied from the context. Coercion is a typical case of such a mismatch. In a coercion sentence, the noun combines with a predicate that typically selects for another semantic type. (4-a) is an example of COERCION, where the verb 'roar' conventionally takes an instance of *human* instead of a *location* as its subject. Consequently, in (4-a), the contextual type (ct) of 'stadium' is *human*, while its lexical type (lt) is *location*, resulting in a clear mismatch. These cases are annotated as COERCION. In our dataset, 81 sentences receive this label.

Mismatch between lt and ct can also arise from other mechanisms, such as metaphor (e.g.(4-b)) and metonymy, though such instances are comparatively rare in our sample. We annotate them under the label OTHER_MISMATCH. Sixteen sentences in the dataset are labeled as OTHER_MISMATCH. When these two mismatch types are discussed together without distinction, we refer to them collec-

[1] For example, many nouns associated to the type *info* tend to be also related to *artifact*; however, this study focuses exclusively on nouns with only one single type.

	matching	coercion	other_mis.	unrestricted
Types	$lt = ct$	$lt \neq ct$	$lt \neq ct$	$ct = \emptyset$
#	1410	81	16	82

Table 1: The meanings of the annotation labels and the number of sentences with the labels

tively as MISMATCH.

(4) a. One moment the **stadium** was roaring and the next, everything went completely silent. (*lt*: *location*, *ct*: *human*)
b. Just because I'm not a social **butterfly** doesn't mean I'm not smart or capable. (*lt*: *animal*, *ct*: *human*)

For all sentences labeled as COERCION or OTHER_MISMATCH, we additionally annotate the contextual type.[2]

Besides the above two cases, there is a third situation that the context is highly general and imposes little selectional restriction on the semantic type of the noun. In such cases we say $ct = \emptyset$. Such contexts are theoretically compatible with almost any types of noun. In example (5), 'linguist' can be replaced by any other types of noun such as 'stadium', 'butterfly', or 'conference' without giving rise to semantic anomaly. These cases are annotated as UNRESTRICTED. Our dataset contains 82 such sentences.

(5) Perhaps you just need a **linguist**. (*lt*: *human*, *ct*: $\emptyset$)

An overview of the annotation labels and their relation with the contextual type is summarized as Table 1.

4 Method

Based on the annotated dataset, we construct graphs in which each node corresponds to an annotated instance of a noun. Edges are established according to the similarity between the embeddings of the instances. For each node, we examine its neighboring nodes with respect to their semantic types and propose quantitative measures to characterize the distribution of neighbor types. By comparing these measures across instances belonging to different sentence categories (i.e. MATCHING, COERCION, OTHER_MISMATCH, UNRESTRICTED), we assess the extent to which the relation between the lexical type and the contextual type of an instance is reflected in the graphs.

4.1 Models

We experiment with the contextualized word embeddings of two language models: BERT (Devlin et al., 2019) (base, uncased) and sense-enhanced BERT (Yavas et al., 2025). Both models are accessed via Hugging Face[3] and the Transformers library (Wolf et al., 2020). Sense-enhanced BERT is a variant of BERT created by fine-tuning in order to incorporate external semantic knowledge into the model embeddings. It is created by fine-tuning BERT on the SemCor corpus (Miller et al., 1993) with WordNet supersense labels. This model is relevant to our study since supersenses are broad semantic categories that correspond well to semantic types, such as *animal*, *location*, and *state*.

We extract the embeddings of the target words in the sentences from the last 4 layers of the models and average them to obtain one embedding per target word instance. Averaging the last four layers has been previously used for semantics-related tasks (Liu et al., 2021; Yavas et al., 2025). If the target word is tokenized into subwords by the model tokenizer, we average the embedding of each subtoken before averaging across layers.

To obtain embeddings without lexical type information, we mask the target word in each sentence using the [MASK] token, as shown in (3-b). We extract the embedding of the [MASK] token and use these masked word embeddings for constructing graphs.

4.2 Graph construction

We conduct a graph-based analysis on the embeddings. Concretely, we construct a directed graph $G = (V, E)$ over the instances in the dataset, where $V = \{v_1, \ldots, v_n\}$ is the set of nodes representing the n instances. Each instance is represented by a contextual word embedding $\mathbf{emb}_i$.

To ensure that edges reflect conceptual similarity between different word types rather than lexical overlap, we only form edges between instances of different target words. A directed edge $(v_i, v_j) \in E$ is formed if v_j is among the k closest neighbors of v_i, determined by the cosine similarity between their embeddings, as in:

$$sim_{ij} = \cos(\mathbf{emb}(v_i), \mathbf{emb}(v_j)),$$

[2]Lexical types are already known from the noun selection process and therefore do not require re-annotation.

[3]Models: *google-bert/bert-base-uncased* and *yavasde/sense-enhanced-bert*. Via: https://huggingface.co/

$E = \{(v_i, v_j) | sim_{ij}$ is among the k highest similarities between v_i and other nodes$\}$

In this paper we set $k = 10$.[4]

As introduced in 4.1, we use four different types of embeddings: BERT, sense-enhanced BERT, masked BERT, masked sense-enhanced BERT. The graphs using these four embeddings are referred to as G_b, G_s, G_{mb} and G_{ms}.

4.3 Neighbor type metrics

We examine the out-neighborhood of each instance, $\mathcal{N}^+(v_i)$, in terms of their types in order to evaluate how well the neighbors reflect the lexical type and the contextual type of the instance. As introduced in 3, in our dataset, for each instance v_i, its lexical type $lt_i \in \mathcal{T}$ and contextual type $ct_i \in \mathcal{T} \cup \emptyset$[5] can be inferred from the annotation. For example, the instance 'stadium' has a lexical type (lt) *location* and a contextual type (ct) *human* in (4-a).

We calculate the distributions of semantic types in the neighborhood of each instance using the Neighbor Type Probability (NTP). For each type $t \in \mathcal{T}$, $NTP(t, v_i)$ measures the proportion of neighbors of v_i that have type t, where t_j denotes the type of neighbor v_j:

$$NTP(t, v_i) = \frac{|\{v_j \in \mathcal{N}^+(v_i) \mid t_j = t\}|}{k}$$

Using NTP, we can determine how much the neighbors of an instance reflect its lexical type lt_i or its contextual type ct_i by calculating $NTP(lt_i, v_i)$ and $NTP(ct_i, v_i)$. We refer to the specific case where NTP is computed for the lexical type lt as the Lexical Neighbor Type Matching Ratio ($NTMR_L = NTP(lt, v)$), and for the contextual type ct as the Contextual Neighbor Type Matching Ratio ($NTMR_C = NTP(ct, v)$).

We propose another metric to measure the diversity of the neighbors in terms of their semantic types: Neighbor Type Entropy (NTE). NTE calculates the entropy of NTP distribution. We define the Neighbor Type Entropy NTE as:

$$NTE(v_i) = -\Sigma_{t \in \mathcal{T}} NTP(t, v_i) \times \log(NTP(t, v_i))$$

For an instance v_i, if most of its neighbors have the same type, the NTE value is relatively low; if the distribution of the types of its neighbors are diverse, the NTE value is higher.

In general, if the graph is organized based on semantic type relations, the neighbors of an instance should be of the same (lexical or contextual) type or taxonomically related semantic types. Moreover, if the graph highlights the lexical information of the instances, we expect most neighbors of an instance v_i to share the lexical type lt_i, i.e. $NTP(lt_i, v_i) \approx 1$. In contrast, if the graph reflects contextual information more than lexical information, we expect $NTP(ct_i, v_i) \approx 1$.

Furthermore, we expect the values of both NTP and NTE to vary across different sentence types (MATCHING, MISMATCH and UNRESTRICTED). These values are obtained by averaging NTP and NTE values of instances belonging to each sentence type. As to the lexical types in MATCHING sentences, since we avoided polysemous uses of the selected nouns in our dataset, the lexical types of the nouns remain stable. Thus, they are expected to be reliably reflected by the lexical types of the neighbors in the graph. More specifically, the $NTMR_L$ of an instance is expected to be high, and the NTE low.

As to MISMATCH sentences, the lexical type of the instance is different from the contextual type. Given that the majority of the instances in our dataset are MATCHING sentences, we expect the types of the neighbors of the instance to be balanced between the two possible options. In this case $NTMR_C$ is likely to be lower and NTE slightly higher. For the instances UNRESTRICTED sentences, the contextual types are unclear, so we expect the contextual types of their neighbors to be more diverse. In this case, the $NTMR_C$ of the contextual type should be low and the NTE should be the highest among all sentence categories.

5 Result analysis

5.1 MATCHING sentences

The Neighbor Type Probability (NTP) is calculated for each instance. Table 2 reports the average $NTMR$ of the instances across different graphs and sentences types. Since G_{mb} and G_{ms} are constructed to capture the information about concextual types instead of lexical types, the $NTMR_L$ on these graphs are significantly lower than on G_b and G_s. Therefore, for MATCHING sentences, our

[4] We set $k = 10$ without systematic evaluation of other values. Since our analysis focuses on relative comparisons across models and sentence types, we expect our conclusions to hold for other values of k.

[5] Theoretically, the contextual type could be a type outside our selected set of type $\mathcal{T}$, but this kind of sentence does not exist in our dataset.

Graph	Sent. type	$NTMR_L$	$NTMR_C$	*other*
G_b	Matching	0.63	—	0.37
	Coercion	0.50	0.17	0.33
	Other_m.	0.54	0.23	0.23
	Unrestr.	0.49	—	0.51
G_{mb}	Matching	0.39	—	0.61
	Coercion	0.18	0.36	0.46
	Other_m.	0.09	0.47	0.44
	Unrestr.	0.14	—	0.86
G_s	Matching	0.81	—	0.19
	Coercion	0.65	0.18	0.17
	Other_m.	0.50	0.41	0.08
	Unrestr.	0.65	—	0.35
G_{ms}	Matching	0.44	—	0.56
	Coercion	0.20	0.40	0.40
	Other_m.	0.11	0.55	0.34
	Unrestr.	0.18	—	0.82

Table 2: Average Lexical Neighbor Type Matching Ratio ($NTMR_L$) and Contextual Neighbor Type Matching Ratio ($NTMR_C$) and the proportion of neighbor types other than lt and ct ($other$) of different sentence types.

Word	G_b	G_s
linguist	80.6	100.0
programmer	48.1	100.0
student	9.3	100.0
genius	54.0	99.3
lady	40.6	99.38
terrorist	50.6	92.6
teenager	88.6	100.0
coward	63.0	81.5
passenger	15.0	43.7
german	86.6	91.6

Table 3: Average Lexical Neighbor Type Matching Ratio ($NTMR_L$) of *human* type words on G_b and G_s

	alien	robot	anim.	arti.	hum.	other
alien	/	0.27	0.28	0	0.39	0.06
robot	0.39	/	0.24	0.34	0	0.03

Table 4: The neighboring words or types of 'alien' and 'robot' and their distributions.

analysis focuses on comparing G_b and G_s.

Fig. 2 presents the average NTP of instances in each lexical type in MATCHING sentences.[6] Both heatmaps indicate that the lexical types of most neighbors coincide with those of the instances themselves. Specifically, on G_b, instances of types *animal*, *activity* and *food* exhibit an average $NTMR_L$ over 80%. This proportion is relatively lower for instances of *human*, *artifact* and *mood*, where fewer than half of their neighbors share the same lexical types. In contrast, on G_s, the average $NTMR_L$ increases for almost all types, most notably for types *artifact* and *human*. These observations suggest that G_s reflects the lexical types of the instances more faithfully than G_b.

A closer investigation of the NTP for *human* instances further illustrates the increase of $NTMR_L$ on G_s. According to Table 3, certain words, such as 'student' and 'passenger', have very low $NTMR_L$ values on G_b. In particular, most neighbors of 'student' instances are of type *location* (predominantly 'classroom' and 'school'), while most neighbors of 'passenger' instances are instances of *artifact* (mainly 'bus' and 'truck'). This indicates that G_b reflects not only semantic type of nouns but also co-occurrence information, whereas on G_s, the semantic type information is more prominently highlighted.

The NTP computed on G_s also reveals the distances between semantic types. As shown in Fig. 2, the NTP between certain pairs of types is relatively higher than between others, indicating their taxonomic proximity. For example, the type *artifact* is close to *food* and *location*; *activity* is close to *process*; *process*, *state* and *mood* form a cluster; and interestingly, *info* is close to *mood*. Although this is less common in linguistic studies on type hierarchies, our result suggests a potential prominence in the distinction of abstract and concrete entities. Although this observation requires further investigation, based on the observed NTP patterns, we can build a new semantic type hierarchy, presented in Fig. 3.

Furthermore, the NTP calculated on G_s can also suggest the types of some non-typical instances. For example, the ontological status of 'alien' and 'robot' is not straightforward. Using our method, according to Table 4, it can be inferred that *alien* is a non-typical member of the type *human*, and *robot* is a non-typical member of *artifact*, and they are close to each other and both similar to the type *animal*.

5.2 UNRESTRICTED sentences

For reasons similar to those observed for MATCHING sentences, the $NTMR_L$ values of instances in UNRESTRICTED sentences are higher when computed with unmasked models and with sense-enhanced models. Compared with MATCHING sentences, however, the $NTMR_L$ values in UNRESTRICTED sentences are lower across all four graphs. The decrease is more significant with

[6] More detailed results are given in the appendix.

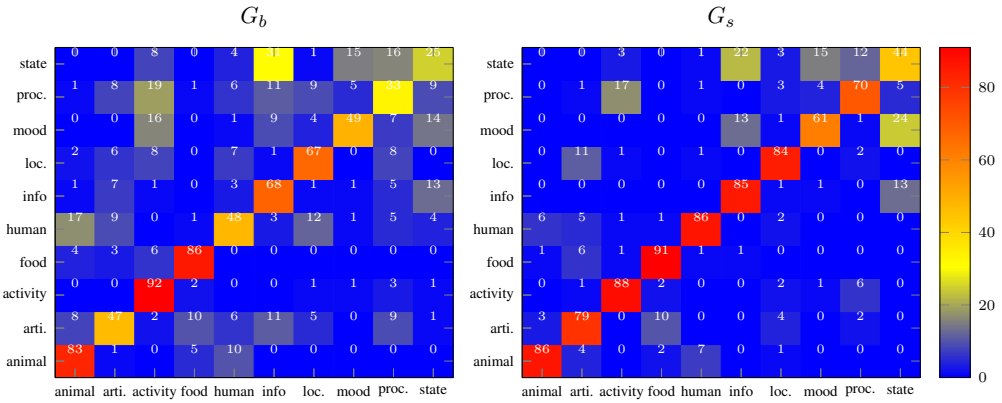

Figure 2: Neighbor Type Ratio (NTP) for MATCHING sentences per semantic type, from the graphs based on BERT (G_b) (left) and sense-enhanced BERT embeddings (G_s) (right). The diagonal grids correspond to the average Lexical Neighbor Type Matching Ratio ($NTMR_L$) for each type.

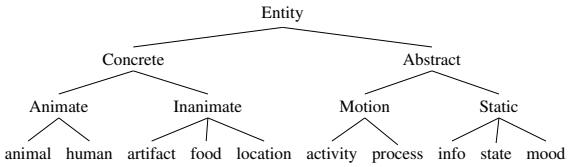

Figure 3: A new type hierarchy induced from the NTP values of the semantic types.

masked models, suggesting that the greater diversity of possible contextual types in UNRESTRICTED sentences is reflected more clearly on G_{mb} and G_{ms}.

The NTE further captures the distinction between MATCHING and UNRESTRICTED sentences. Across all four graphs, the diversity of neighbor type distributions is significantly higher for UNRESTRICTED sentences than for MATCHING sentences.

5.3 MISMATCH sentences

The results in Table 2 show that across all graphs, the $NTMR_L$ values of MISMATCH sentences (both COERCION and OTHER_MISMATCH) are lower than those of MATCHING sentences. This indicates that the different contextual types exert a noticeable influence on the organization of the graphs.

As discussed in the previous section, the graphs G_b and G_s largely reflect the information about lexical types. On these graphs, most neighbors of the instances in COERCION sentences share the same lexical types instead of the contextual types. This finding seems to suggest that instances in COERCION sentences retain their lexical types and do not undergo a type shift. A more in-depth research is needed to confirm this observation.

With respect to $NTMR_C$, the graphs constructed with masked models produce higher values than the others. As noted above, in the graphs constructed with unmasked models, especially with sense-enhanced bert, the neighbors of an instance usually have the same lexical type. Consequently, the contextual types of MISMATCH sentences are underrepresented in these graphs, resulting in $NTMR_C < NTMR_L$.

Table 5 reveals that no significant difference is found between COERCION and OTHER_MISMATCH sentences. Therefore, due to the higher number of instances in the dataset, we mainly focus the analysis of the result on COERCION sentences.

A comparison of NTE between COERCION and UNRESTRICTED sentences reveals a pattern distinct from that observed between MATCHING and COERCION sentences. On G_{mb} and G_{ms}, the NTE of COERCION sentences is significantly lower than that of UNRESTRICTED sentences, whereas on G_b and G_s, no significant difference is found between the two sentence types. This implies that contextual type information is reflected on G_{mb} and G_{ms} to some extent.

For G_{mb} and G_{ms}, although $NTMR_C$ is higher than $NTMR_L$, the average value remains lower than 0.5, indicating that the information about contextual types is not highly prominent in these graphs. This observation is further supported by the analysis of NTE. According to Table 5, the NTE of instances in MATCHING sentences is significantly lower than the NTE of COERCION sentences on G_b and G_s, but not on G_{mb} and G_{ms}.

G	mat.	coer.	other.	unres.	Comparison
G_b	0.96	1.14	0.99	1.16	mat. $<^{**}$ coer. mat. $<^{**}$ unres. coer. $\not<$ unres. coer. $\neq$ other.
G_{mb}	1.71	1.67	1.64	1.96	mat. $\not<$ coer. mat. $<^{***}$ unres. coer. $<^{*}$ unres. coer. $\neq$ other.
G_s	0.61	0.85	0.65	0.83	mat. $<^{***}$ coer. mat. $<^{***}$ unres. coer. $\not<$ unres. coercion $\neq$ other.
G_{ms}	1.45	1.49	1.42	1.77	mat. $\not<$ coer. mat. $<^{***}$ unres. coer. $<^{**}$ unres. coer. $\neq$ other.

Table 5: Mean Neighbor Type Entropy (NTE) of different sentence types and the statistical comparisons between pairs (Mann-Whitney U test) . $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. $\not<$ indicates the hypothesis was not confirmed. The sentence types MATCHING, COERCION, OTHER_MISMATCH and UNRESTRICTED are shortened as MAT., COER., OTHER. and UNRES. respectively.

This suggests that G_{mb} and G_{ms} only partially reflect the information about contextual types. We observe that the graphs sometimes capture other information beyond semantic types, such as morphological clues and co-occurrence patterns. For example, on G_{mb} and G_{ms}, the neighbors of the instances of 'elephant' are often instances of 'omelette' and 'explosion', possibly due to the fact that they are all followed by the article 'an'.

Moreover, the information of lexical types is not completely eliminated from G_{mb} and G_{ms}. In these two graphs, the NTP of the lexical types are still higher than the NTP of other types. This suggests that the lexical information may persist in COERCION sentences. For example, in example (2), although 'finish' strongly indicates a contextual type *activity*, the other words in the sentence, such as 'gulp', still implies the existence of an *artifact* bottle.

In summary, the three types of sentences and their relations with NTE values across the four graphs are summarized as Table 6. This can potentially help in detecting coercion instances.

	NTE on G_b/G_s	NTE on G_{mb}/G_{ms}
matching	low	low
coercion	high	low
unrestr.	high	high

Table 6: The relative numerical relations of Neighbor Type Entropy (NTE) values among different types of sentences.

6 Conclusive remarks

This paper has presented a graph-based analysis of the contextualized word embeddings of noun instances, examining how semantic type information is reflected across different sentence types. We selected nouns from ten semantic types, extracted corpus instances for each noun, and annotated them with lexical types and contextual types. Instances in which the lexical type coincides with the contextual type are labeled as MATCHING; instances where the two diverge are labeled as MISMATCH, further classified as COERCION or OTHER_MISMATCH; and instances occurring in uninformative contexts are labeled as UNRESTRICTED.

Using these annotated instances, we constructed graphs based on the similarity between their embeddings and proposed two graph-based metrics in order to analyze the neighbors of instances based on their semantic types: Neighbor Type Probability (NTP) and Neighbor Type Entropy (NTE). Our results demonstrate that graphs constructed with sense-enhanced embedding reflect information about semantic types better than BERT embeddings. Lexical type information is reliably reflected in graphs constructed on sense-enhanced embeddings, and contextual type information is also partially reflected in graphs constructed on masked embeddings. MATCHING, MISMATCH, and UNRESTRICTED sentences can be distinguished from one another through the comparison of NTP and NTE values across the graphs.

Our method also has a number of potential future applications. On the linguistic side, the distinction between coercion and other mechanisms underlying lexical-contextual type mismatch can be further explored, given sufficient number of instances with metaphor, metonymy or other kinds of mismatch. A preliminary new type hierarchy has been implied from our method, and with more selection of nouns and, preferably, larger unannotated data, semantic types and their hierarchical organization can possibly be induced automatically, and MISMATCH sentences can possibly be consistently detected.

On the computational side, the effectiveness of our graph-based method suggests that it can be used to study the mechanisms by which the lexical and contextual type information are encoded and how they interact with each other.

References

Nicholas Asher. 2011. *Lexical meaning in context: A web of words*. Cambridge University Press.

Nicholas Asher. 2015. Types, meanings and coercions in lexical semantics. *Lingua*, 157:66–82.

Nicholas Asher, Tim Van de Cruys, Antoine Bride, and Márta Abrusán. 2016. Integrating type theory and distributional semantics: A case study on adjective–noun compositions. *Computational Linguistics*, 42(4):703–725.

Long Chen, Laura Kallmeyer, and Rainer Osswald. 2022. A frame-based model of inherent polysemy, copredication and argument coercion. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 58–67, Taipei, Taiwan. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT press.

Frederick Gietz and Barend Beekhuizen. 2022. Remodelling complement coercion interpretation. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 158–170, online. Association for Computational Linguistics.

Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2021. MirrorWiC: On eliciting word-in-context representations from pretrained language models. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 562–574, Online. Association for Computational Linguistics.

Zhaohui Luo. 2012. Formal semantics in modern type theories with coercive subtyping. *Linguistics and Philosophy*, 35(6):491–513.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Richard Montague. 1970. Universal grammar. *Theoria*, 36(3):373–398.

James Pustejovsky. 1993. Type coercion and lexical selection. In *Semantics and the Lexicon*, pages 73–94. Springer.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

Matteo Radaelli, Emmanuele Chersoni, Alessandro Lenci, and Giosuè Baggio. 2025a. Compositionality and event retrieval in complement coercion: A study of language models in a low-resource setting. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 469–480, Vienna, Austria. Association for Computational Linguistics.

Matteo Radaelli, Emmanuele Chersoni, Alessandro Lenci, and Giosuè Baggio. 2025b. Context effects on the interpretation of complement coercion: A comparative study with language models in Norwegian. In *Proceedings of the 16th International Conference on Computational Semantics*, pages 78–88, Düsseldorf, Germany. Association for Computational Linguistics.

Giulia Rambelli, Emmanuele Chersoni, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2020. Comparing probabilistic, distributional and transformer-based models on logical metonymy interpretation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 224–234, Suzhou, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Deniz Ekin Yavas, Timothée Bernard, Benoit Crabbé, and Laura Kallmeyer. 2025. On the relation between fine-tuning, topological properties, and task performance in sense-enhanced embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23610–23625, Vienna, Austria. Association for Computational Linguistics.

Deniz Ekin Yavas, Laura Kallmeyer, Rainer Osswald, Elisabetta Jezek, Marta Ricchiardi, and Long Chen. 2023. Identifying semantic argument types in predication and copredication contexts: A zero-shot cross-lingual approach. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 310–320, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bingyang Ye, Jingxuan Tu, Elisabetta Jezek, and James Pustejovsky. 2022. Interpreting logical metonymy through dense paraphrasing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Mengjie Zhao, Philipp Dufter, Yadollah Yaghoobzadeh, and Hinrich Schütze. 2020. Quantifying the contextualization of word representations with semantic class

probing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1219–1234, Online. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Appendix

Semantic Type	Word	G_b	G_s
info	message	63.75	100.00
	information	88.75	90.00
	poem	13.13	100.00
	rumor	88.75	100.00
	data	81.33	90.00
	news	100.00	100.00
	idea	93.75	96.88
	concept	89.38	100.00
	science	23.75	61.25
	knowledge	78.13	67.50
	problem	10.63	20.63
	secret	81.88	99.38
mood	happiness	31.88	70.00
	empathy	76.88	66.88
	resentment	86.25	97.50
	preference	46.25	75.00
	attention	23.13	13.13
	mood	43.75	65.63
animal	elephant	87.33	99.33
	dinosaur	97.50	96.88
	pig	82.00	79.33
	cat	100.00	98.46
	cow	67.33	84.00
	butterfly	100.00	100.00
	penguin	100.00	93.85
	pigeon	99.33	94.00
	mammal	99.38	100.00
	bug	90.91	96.36
location	cave	99.38	98.75
	classroom	58.75	87.50
	stadium	37.50	93.75
	school	88.75	89.38
	river	88.75	90.00
	road	43.75	22.50
	valley	100.00	99.38
	desert	95.00	100.00
	sky	11.88	80.63
	park	60.00	95.00
state	death	24.38	24.38
	existence	26.88	79.38
	chaos	21.88	46.25
	difficulty	21.25	13.75
	strength	63.75	91.25
	beginning	16.88	53.13
	intelligence	7.50	20.63

Table 7: The average $NTMR_L$ value of each noun per semantic type on G_b and G_s

Semantic Type	Word	G_b	G_s
activity	picnic	88.00	91.33
	barbecue	71.88	61.88
	banquet	90.63	90.63
	celebration	83.75	75.00
	festival	98.13	97.50
	carnival	96.00	94.00
	conference	99.38	90.00
	meeting	98.75	97.50
	parade	86.67	77.33
	party	90.63	85.00
process	explosion	56.25	99.38
	accident	74.38	85.63
	experiment	12.50	78.13
	installation	28.13	90.63
	interruption	15.00	87.50
	storm	20.63	68.13
	sleep	0.00	12.67
	dance	16.25	17.50
	attack	65.00	95.63
	performance	23.13	53.75
food	pizza	91.88	88.75
	kebab	100.00	100.00
	rice	97.50	100.00
	omelette	100.00	100.00
	pork	86.88	95.63
	beef	100.00	100.00
	bread	98.75	100.00
	cake	86.88	100.00
	candy	73.57	92.86
	cigarette	16.88	10.63
	apple	93.57	100.00
	carrot	84.00	96.67
	potato	100.00	100.00
artifact	bus	90.00	100.00
	truck	68.75	100.00
	bicycle	81.88	100.00
	rocket	12.14	58.57
	guitar	43.75	93.75
	sculpture	1.25	71.25
	computer	1.88	72.50
	shirt	80.63	97.50
	coin	25.63	48.13
	cable	43.75	85.63
	button	77.50	99.38
	bottle	69.38	68.13
human	linguist	80.63	100.00
	programmer	48.13	100.00
	student	9.38	100.00
	genius	54.00	99.33
	lady	40.63	99.38
	terrorist	50.67	92.67
	teenager	88.67	100.00
	coward	63.08	81.54
	passenger	15.00	43.75
	german	86.67	91.67

Table 8: The average $NTMR_L$ value of each noun per semantic type on G_b and G_s