CALE : Concept-Aligned Embeddings for Both Within-Lemma and Inter-Lemma Sense Differentiation

Bastien Liétard[1] **and Gabriel Loiseau**[1,2]**,**
[1]University of Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
[2]Hornetsecurity, Hem, France
first_name.last_name@inria.fr
Authors contributed equally.

Abstract

Lexical semantics is concerned with both the multiple senses a word can adopt in different contexts, and the semantic relations that exist between meanings of different words. To investigate them, Contextualized Language Models are a valuable tool that provides context-sensitive representations that can be used to investigate lexical meaning. Recent works like XL-LEXEME have leveraged the task of Word-in-Context to fine-tune them to get more semantically accurate representations, but Word-in-Context only compares occurrences of the same lemma, limiting the range of captured information. In this paper, we propose an extension, Concept Differentiation, to include inter-words scenarios. We provide a dataset for this task, derived from SemCor data. Then we fine-tune several representation models on this dataset. We call these models Concept-Aligned Embeddings (CALE). By challenging our models and other models on various lexical semantic tasks, we demonstrate that the proposed models provide efficient multi-purpose representations of lexical meaning that reach best performances in our experiments. We also show that CALE's fine-tuning brings valuable changes to the spatial organization of embeddings.

1 Introduction

Research in computational lexical semantics has relied on contextualized embeddings to study word meaning in context (Neidlein et al., 2020; Chronis and Erk, 2020; Apidianaki and Garí Soler, 2021; Yu and Xu, 2023; Li et al., 2024), with applications extending beyond traditional tasks, such as in debate modeling (Garí Soler et al., 2023) and political discourse analysis (Boholm et al., 2024). However, most approaches use pre-trained models like BERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020) without fine-tuning specifically for lexical semantics, due to limited annotated data.

While several models are fine-tuned for the Word-in-Context (WiC) task (Pilehvar and Camacho-Collados, 2019; Liu et al., 2021; Cassotti et al., 2023; Mosolova et al., 2024), WiC in its traditional definition only captures within-word meaning variation (i.e. the word's senses) and ignores inter-word semantic relations. Consequently, models trained on WiC may struggle to capture the broader structure of lexical meaning.

To reach a higher perspective, the focus must shift from word senses to *semantic concepts*. Throughout this paper, we distinguish senses and concepts as follows: a word sense is a conventionalized way to use a particular lemma in relevant contexts, it characterizes the use of a specific word for an intended meaning; concepts are mental representations of categories of objects, events, acts and ideas, and we use words to refer to them (Murphy, 2002). In context, the meaning of a word is the concept it refers to, and a word sense is a pointer to a concept. Our concepts are equivalent to Wikipedia entries, or WordNet synsets (Miller, 1995) as made in Liétard et al. (2024).

In this paper, we are interested in the use of contextualized models to obtain multi-purpose (as opposed to "task-specific") vector representation of word meaning. Word occurrences that instantiate the same concept in respective contexts should have similar representations, while occurrences of words that refer to different concepts should have dissimilar embeddings. A model with this desired property would be useful in studies of the mapping between word forms and concepts, like Haspelmath (2023) for instance, or in any computational lexical semantic applications mentioned above.

We focus on *synonymy* (different words having the same meaning, instantiating the same concept) and *polysemy* (a single word that can refer to multiple concepts) and we leave to future work the integration of other relations such as hyponymy.

First, we introduce Concept Differentiation, a task that generalizes WiC to compare contextualized meanings of both same-lemma and cross-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1086–1100

lemma word pairs. Given two occurrences, the task is to determine whether they instantiate the same semantic concept. Like WiC is to Word Sense Induction, Concept Differentiation is a binary classification form of Concept or Frame Induction (Liétard et al., 2024; QasemiZadeh et al., 2019).

To support this task, we construct a new dataset: SPCD (Semcor Pairs for Concept Differentiation), drawn from SemCor (Miller, 1995), including both intra- and inter-lemma pairs. This enables evaluation of lexical meaning representations at both local (sense-level) and global (concept-level) scales.

Then, we develop CALE (Concept-Aligned Embeddings), a family of token-level models fine-tuned for Concept Differentiation using SPCD. We compare them to their base pre-trained counterparts and to XL-LEXEME (Cassotti et al., 2023), a strong multilingual model trained on WiC-style data.

CALE achieve strong performance across evaluations: 79.3 balanced accuracy on Concept Differentiation, as well as top scores on Lexical Semantic Change Detection (a sense-level task) and in-context Lexical Similarity (a concept-level task). CALE models even match or outperform XL-LEXEME in non-English benchmarks, despite monolingual fine-tuning, demonstrating the strength of their fine-tuning objective and the base model XLM-R's cross-lingual capabilities.

Finally, analysis of the resulting embedding spaces reveals that CALE fine-tuning enables the shift from lemma-centric to concept-centric representations. We also show that CALE best reflects the conceptual structure of our reference lexical resources WordNet, with a correlation to similarities in WordNet above .50. We release both the SPCD dataset and the CALE models to support further research on token-level lexical meaning representations.

2 Related Work

2.1 Lexical Semantic Tasks

A variety of NLP tasks have explored the mapping between words and meanings, especially regarding polysemy and synonymy (or similarity in meaning in general). Word Sense Disambiguation (WSD) maps word occurrences to predefined senses (Navigli, 2009; Bevilacqua et al., 2021), while its unsupervised counterpart, Word Sense Induction (WSI), clusters usages into latent sense groupings (Manandhar et al., 2010; Jurgens and Klapaftis, 2013). The Word-in-Context (WiC) task refines this by classifying whether two instances of a word share the same sense (Pilehvar and Camacho-Collados, 2019; Martelli et al., 2021). Lexical Semantic Change Detection (LSCD) extends WSI temporally, aiming to detect or rank meaning shifts across time periods (Schlechtweg et al., 2020; Zamora-Reina et al., 2022). While these tasks focus on polysemy and intra-word variation, they do not address inter-word semantic similarities. Tasks such as Lexical Substitution (McCarthy and Navigli, 2007; Zhou et al., 2019) and Lexical Similarity (Hill et al., 2015; Vulić et al., 2020; Huang et al., 2012; Armendariz et al., 2020) examine inter-word relations, centering on synonymy and contextual interchangeability. Substitution tasks often involve ranking appropriate alternatives in context (Erk and Padó, 2008; Kremer et al., 2014).

More recently, Concept Induction has been proposed to unify word-level and cross-word perspectives by clustering usages of multiple target words into semantic concepts (Liétard et al., 2024). Similarly, Frame Induction clusters word usages (typically verbs) into broader event frames with their associated arguments (QasemiZadeh et al., 2019; Yamada et al., 2021; Mosolova et al., 2024). These tasks aim to capture higher-order abstractions that go beyond individual word senses or pairwise similarity.

2.2 Embeddings for Semantic Representations

Contextualized word embeddings from models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) generate dynamic representations based on usage in context, enabling better semantic modeling than earlier static or sparse methods. However, pooling token-level embeddings into meaningful sentence-level vectors remains challenging due to issues like anisotropy in the embedding space[1] (Reimers and Gurevych, 2019). To improve semantic similarity and retrieval, Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) introduced a Siamese training architecture, followed by advances using contrastive learning (Gao et al., 2021) and multi-task objectives (Zhang et al., 2022) to enhance cross-task generalization. Such models have been adapted for lexical semantics. Mosolova et al. (2024) fine-tune language models using Wiktionary supervision to improve con-

[1] i.e., the tendency for embeddings to occupy a narrow cone, leading to high cosine similarity for even unrelated items (Gao et al., 2019).

textualized token embeddings for tasks like Word-in-Context. Similarly, the XL-LEXEME model (Cassotti et al., 2023), based on XLM-RoBERTa (Conneau et al., 2020), extends SBERT for multilingual word meaning tasks, showing robustness on LSCD benchmarks and supporting cross-lingual, context-sensitive embedding spaces.

Other efforts fine-tune BERT for word-level classification tasks (Garí Soler and Apidianaki, 2020), though these cross-encoder models often lack flexibility for single-occurrence representation which represents a limitation for tasks requiring generalized concept modeling. Embedding-based models have also proven effective in capturing non-literal or stylistic nuances, as shown by applications in authorship verification (Rivera-Soto et al., 2021). Additionally, SenseBERT (Levine et al., 2020) enhances contextual embeddings by incorporating sense information from WordNet during pretraining, improving performance on sense-sensitive tasks and highlighting the benefit of explicitly modeling lexical semantics in transformer-based representations. Collectively, these developments demonstrate the adaptability of sentence-transformer architectures for fine-grained semantic representations.

3 Concept Differentiation

Our primary goal is to build a model able to accurately represent semantic information in order to distinguish concepts in context, from which we can obtain concept-aligned embeddings. To train such model, we rely on a supervised task that captures both polysemy and synonymy and propose Concept Differentiation, a supervised alternative to Concept Induction defined in Liétard et al. (2024).

3.1 Task Definition

We define Concept Differentiation as a binary classification task where the goal is to determine whether two word usages represent the same underlying concept, regardless of whether they come from the same lemma or different lemmas. To illustrate, consider the following three phrases:

- A. "the boy could easily distinguish the different note values"
- B. "the patient's ability to recognize forms and shapes"
- C. "the government had refused to recognize their autonomy and existence as a state"

The verbs *distinguish* in A and *recognize* in B both express the concept of DISCERN — to tell things apart — so this pair is labeled 1 (same concept). In contrast, while B and C both use *recognize*, only C refers to ACKNOWLEDGE, so that pair is labeled 0 (different concepts). This task is broader than traditional Word-in-Context, which is limited to same-lemma comparisons. Here, we extend the comparison to any pair of word usages — same lemma or not — capturing finer and more flexible distinctions of meaning. It can also be seen as a binary reformulation of the Concept Induction task, where instead of clustering occurrences into concept clusters, we ask if two usages share a common concept. This setup allows us to train models that learn to compare usages in context, effectively capturing subtle semantic distinctions or consistencies between different instances of target words across the corpus.

3.2 Dataset for Concept Differentiation

In the absence of a dedicated dataset for this task, we construct a new dataset using the SemCor corpus, using its WordNet synsets annotations (Miller, 1995; Fellbaum, 1998) as concept labels. For each lemma with enough labeled instances, we retrieve its occurrences following the procedure detailed in Appendix A. We gather 70.3k occurrences for 1902 target lemmas, covering 5899 concepts. To create labeled training pairs, we sample pairs of word occurrences (o_j, o_k) and assign a label of 1 if both instances are annotated with the same concept, and 0 otherwise, regardless of whether they are occurrences of the same target lemma or not.

To rigorously assess generalization, we enforce a strict separation between training, validation, and test sets by first partitioning the set of concepts and the set of target words into disjoint subsets for each split. Specifically, we randomly sample subsets of concepts and lemmas to be assigned exclusively to the validation or test split (5% of concepts / lemmas in validation, 10% in test). Once these partitions are defined, we extract all word occurrences from the SemCor corpus that are annotated with the selected concepts and lemmas, and include them in the corresponding data split. This ensures a more rigorous evaluation of a model's ability to generalize to entirely unseen meanings and lexical items, preventing models from relying on memorized associations between words and concepts. In each split separately, we create pairs by attempting to match each occurrence to 4 others, one for each

of the following categories if possible: same concept and same lemma (SC/SL), same concept but different lemma (SC/DL, only if we can find an occurrence of another lemma for the concept), different concept but same lemma (DC/SL, only if the word is polysemous) and different concept and different lemma (DC/DL, unrelated pairs). The dataset contains 44k pairs (from the 14.3k reserved occurrences) in the *test* split, 20k (from the 6.5k reserved occurrences) in the *validation* split, and 156k (from the 49.6k remaining occurrences) in the *train* split. In all splits, the resulting proportion of label '1' (same concept) is around 40%.

The resulting dataset, later referred as "SemCor Pairs for Concept Differentiation" (SPCD), supports research in semantic representation and disambiguation and is made publicly available to facilitate future benchmarking and exploration[2]. Further description as well as more details about the extraction process can be found in Appendix A.

4 Concept-Aligned Embeddings

We use our resulting pair classification dataset to train a representation model. The goal is to learn a function f mapping an occurrence o_i to a vector representation e_i in a vector space in which the embeddings of two occurrences of the same concept have higher cosine similarity than embeddings of two occurrences of a different concept. To fit f, we adopt a Siamese representation learning architecture, inspired from SentenceBERT. Concretely, our model consists of a pre-trained contextualized language model followed by a pooling layer that produces a fixed-dimensional vector for an input sequence. During training, two identical copies of this encoder sharing the same parameters are used to process two occurrences as a training pair. The parameters of the encoder are updated jointly, ensuring that both occurrences are embedded in the same representation space.

Each training step considers a batch of occurrence pairs (o_i, o_j) with an associated label $y_{ij} \in \{0, 1\}$, where $y_{ij} = 1$ indicates that both occurrences share the same semantic concept, and $y_{ij} = 0$ otherwise. Occurrence of the pair are encoded into vector representations (e_i, e_j) by the Siamese models (two instances of the same model with shared weights during training). We then compute the cosine similarity between the two embeddings $\text{sim}_{ij} = \cos(e_i, e_j)$. The training objective is to maximize similarity for positive pairs ($y_{ij} = 1$) and minimize it for negative pairs ($y_{ij} = 0$). We fit the model to minimize a contrastive loss (Hadsell et al., 2006) defined as:

$$\frac{1}{2}\left[y \cdot \text{sim}_{ij}^2 + (1-y) \cdot \max(0, m - \text{sim}_{ij})^2\right]$$

where m defines the minimum distance that should exist between dissimilar (negative) pairs in the embedding space, known as the margin.

This contrastive training encourages the model to organize the embedding space such that occurrences sharing the same concept are pulled closer together, while unrelated occurrences are pushed apart. Fine-tuning under this objective enables the model to produce embeddings that better reflect conceptual similarity across varied lexical and contextual realizations.

An occurrence o_i is a sequence of words $(t_1, \ldots, t_n)$ with a target word w in position k ($t_k = w$). To be used as input for the model, the occurrence is transformed as follows:

$$\underbrace{t_1, \ldots, t_{k-1}}_{\text{left context}}, \texttt{<t>}, w, \texttt{</t>}, \underbrace{t_{k+1} \ldots, t_n}_{\text{right context}},$$

using special tags <t></t> to delimit the target word in the sentence. During training, the model will learn that the contrast must be made based on the meaning of words between the delimiters. In doing so, we ensure the embedding corresponding to the full sequence is the representation of $o_{w,i}$ and distinct from the embedding of another target word occurring in the same sentence, without needing to account for potential subwords that may result from the model's tokenizer. This method, also used by Cassotti et al. (2023), makes the model easy to use in practice at inference time, as we only have to add/move the delimiters to shift the model's focus from one word to another in the sentence.

We refer to this fine-tuning approach and the resulting models as **CALE** (Concept Aligned Embeddings) and adopt this terminology throughout the paper. Our models are made publicly available through HuggingFace[3]. We limited the scope of our study to single-word targets, but our framework could be applied to capture the meaning of Multi-Word Expressions in future work.

[2] https://hf.co/datasets/gabrielloiseau/CALE-SPCD/

[3] [Hyperlink to HF-models]

5 Model Evaluation

In this section, we report the performances of different models, including models with CALE fine-tuning, for the Concept Differentiation task on the test split of the SPCD dataset, as well as other external lexical semantic tasks, namely Lexical Semantic Change Detection and In-context Lexical Similarity.

Compared Models. XLM-R (Conneau et al., 2020) is based on the RoBERTa Large architecture and pretrained on cross-lingual data. XL-LEXEME (Cassotti et al., 2023) is a XLM-R model fine-tuned for multilingual WiC. Compared to our models, XL-LEXEME's fine-tuning was rather data-hungry, as three Word-in-Context datasets were used. XL-LEXEME established a new State-of-the-Art in some semantic tasks and is therefore our main reference point. Regrettably, models from Garí Soler and Apidianaki (2020) and Mosolova et al. (2024) are not publicly available, preventing us to include them in our experiments without re-implementing the whole dataset curation process and fine-tuning method. Our CALE models generate embeddings in a similar fashion than XL-LEXEME, but they are fine-tuned on Concept Differentiation, as described earlier. We experiment with several base model: XLM-R (to compare with XL-LEXEME) XL-LEXEME (to observe the impact of double fine-tuning and of XL-LEXEME's fine-tuning multilingually) and a monolingual English ModernBERT (Warner et al., 2024), a more-recent State-of-The-Art actualization of the BERT model, evaluated only in English datasets in order to discuss the potential difference between language specialization and pre-training multilingually.

Embedding extraction. For XLM-R, embeddings of word occurrences are obtained by averaging first over a specified range of layers, and then over sub-word tokens that compose the target word of the occurrence. XL-LEXEME and CALE models are SentenceBERT models, meaning that a single embedding is produced per occurrence using a final pooling layer that averages the last layer embeddings of all tokens in the sequence. In that case, the delimiters around the target word ensure that the representation is specific to that target.

Hyperparameters and metrics. For XLM-R, we test three layer ranges. The `mid` layer set corresponds to intermediate-high layers (14 to 17), following the findings of Chronis and Erk (2020); the `last` set (layers 21 to 24) is in line with a large number of works in lexical semantic tasks using the average over the last four layers as input; and `first` (layers 1 to 4) is included for comprehensiveness. For all models and across all experiments, cosine distance is used as the standard metric of the representational space to compare two embeddings.

	All	SL	DL
1L1C (baseline)	65.1	50.0	50.0
XLM-R (first)	60.8	54.2	50.0
XLM-R (mid)	70.9	59.1	68.4
XLM-R (last)	69.1	59.6	63.5
XL-LEXEME	76.7	62.4	82.7
CALE (XLM-R)	**79.3**	65.9	**86.4**
CALE (MBERT)	78.7	66.6	83.5
CALE (XL-LEX)	79.2	**67.2**	84.4

Table 1: Balanced Accuracies on test pairs from SPCD. "SL" (resp. "DL") to pairs of occurrences of the same lemma (resp. of different lemmas).

5.1 Concept Differentiation Evaluation

In Table 1 we present the results for the Concept Differentiation task on the test split of the SPCD datasets.

Threshold-based classifiers. For each candidate model, the classification process for a given pair of occurrences is the following: we compute the embeddings for each occurrence using the candidate model, then compare the cosine distance between them to a threshold to label the pair: 1 (same concept) if the cosine distance below the threshold, 0 (different concepts) otherwise. The threshold is chosen to maximize accuracy on the train and validation splits.

Baseline. To provide a baseline system, we label a pair as 1 if and only if the two occurrences are from the same lemma, and 0 otherwise. We refer to this baseline as `1L1C` for "1 Lemma 1 Concept".

Evaluation metrics. We use Balanced Accuracy (the average of the Recalls from both classes) as the classes are not exactly balanced. We also tried F1 scores and the observed tendencies are the same, thus we only report them in Appendix C, with also a more complete overview of accuracies per class.

Discussion. Our first observation is that the 1L1C baseline reaches high accuracy, providing a challenging baseline. This is in line with prior obser-

vations in the context of clustering (Liétard et al., 2024). This indicates that the mapping between words and concepts in SemCor may resemble a one-to-one mapping with some deviations, and as such a model for Concept Differentiation would benefit from not ignoring whether two occurrences are of the same lemma or not. We also observe that XLM-R performs best using late-intermediate layers than other layers and outperforms the baseline verifying findings of Chronis and Erk (2020) that these layers are best for tasks related to semantic similarity.

We also find that XL-LEXEME largely outperforms XLM-R even when presented pairs of Different Lemmas, meaning that by learning to differentiate senses of the same lemma, it captured more global information about relations between different words.

We find that CALE models outperforms XL-LEXEME. It is expected to some extent, as Concept Differentiation is the task CALE has been fine-tuned for. We also observe that Monolingual CALE (MBERT) is not better than the multilingual ones, underlining the advantage of multilingual pre-training.

The CALE model based on XL-LEXEME is globally on-par with the one based on XLM-R (79.3 and 79.2 balanced accuracies). The former is slightly better in the Same-Lemma setting (i.e. better at distinguishing senses) than the latter, but slightly worse in the Different-Lemma setting (i.e. worse at identifying synonyms), as expected because XL-LEXEME has been trained strictly in the Same-Lemma setting and therefore likely provides benefit in this context only.

5.2 Lexical Semantic Change Detection.

To assess models' ability to capture meaning variation within lemmas across contexts, we evaluate on the Lexical Semantic Change Detection (LSCD) task, following "Subtask 2" of Schlechtweg et al. (2020). We use Diachronic Word Usage Graphs (DWUGs) datasets across six languages: English, German, Latin, Swedish, Spanish, and Italian (Schlechtweg et al., 2020; Zamora-Reina et al., 2022; Cassotti et al., 2024)[4]. Each dataset contains target words with occurrences sampled from two time periods (T_1, T_2), and the task is to assign a change score reflecting how much a word's meaning has shifted. LSCD is a intra-lemma task: only usages of the same word are compared and semantic change is assessed by measuring differences in the word's senses between time periods. Performance is measured by Spearman correlation between predicted and gold change scores. Note that Cassotti et al. (2023) also evaluated their model XL-LEXEME on these benchmarks and established new State-of-the-Art scores in English, German and Swedish.

[4] https://www.ims.uni-stuttgart.de/en/research/resources/experiment-data/wugs/

	EN	DE	ES	IT	LA	SV
# targets	46	50	100	26	40	44
XLM-R (first)	.22	.14	.08	.18	.00*	.10
XLM-R (mid)	.57	.54	.49	.32	**.24***	.58
XLM-R (last)	.56	.34	.53	.27	.00*	.52
XL-LEXEME	**.79**	.80*	.63*	.51*	.14*	.85*
CALE (XLM-R)	.71*	.78*	.65*	**.75**	.11*	.83*
CALE (MBERT)	.66*	-	-	-	-	-
CALE (XL-LEX)	.73*	**.83**	**.68**	.54	.20*	**.86**

Table 2: Spearman Correlation between model Average Pairwise Divergence (APD) and gold change scores in DWUG datasets. The highest score for each dataset is in bold, and other scores that are not significantly different from it are marked with *.

We use the Average Pairwise Divergence (APD) measure from Kutuzov and Giulianelli (2020) to compute change scores, relying on cosine distances between embeddings from T_1 and T_2:

$$\mathrm{APD}(w) = \frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} d(e^1_{w,i}, e^2_{w,j})$$

with d the cosine distance operator, $e^k_{w,i}$ the embedding of the i-th occurrence of word w at time T_k, for a target word w with n occurrences at T_1 and m occurrences at T_2. APD is generally accepted as standard for Lexical Semantic Change (see the survey study of Periti and Montanelli (2024)) and was used by Cassotti et al. (2023) to evaluate XL-LEXEME on the same task on DWUGs datasets. In Appendix D we tried PRT, another measure from Kutuzov and Giulianelli (2020).

We report APD-based results in Table 2. We use Steiger's Z-test on dependent correlations to determine whether differences to the best-performing model are significant or not (Steiger, 1980).

Results show that CALE fine-tuning consistently improves over base XLM-R in all languages, even though it was trained on English data only, showcasing the cross-lingual transferability of XLM-R. CALE also matches or slightly outperforms XL-LEXEME in 5 out of 6 languages, despite XL-LEXEME being fine-tuned on larger multilingual

Context	Similarity
[...] As J.T. put it, if Kaz can't follow orders, he might as well learn to give them.	2.88
[...] Caleb returns to the table to let Kirsten know that he will give his support to Julie at the board meeting. [...] but for her marriage's sake, she will put her support to Sandy.	7.83

Table 3: Example entry of CoSimLex context pairs with the mean similarity rating between target words (underlined).

datasets. As in prior work, Latin proves challenging across all models.

While XL-LEXEME is fine-tuned for Word-in-Context (a lemma-centric task closely related to LSCD), CALE is optimized for Concept Differentiation, a broader task involving both intra- and inter-lemma comparisons. One might expect this to hinder LSCD performance, yet results suggest otherwise: CALE (XL-LEXEME) slightly improves over XL-LEXEME alone, suggesting that Concept Differentiation enhances a model's ability to detect semantic change. Due to the limited number of target words, differences are not statistically significant, but findings indicate that CALE fine-tuning does not degrade LSCD performance.

5.3 Contextual Lexical Similarity.

We now evaluate candidate models in the cross-lemma setting.

Dataset. To assess our models' ability to represent conceptual similarity of two distinct words in a given context, we use the CoSimLex dataset (Armendariz et al., 2020) for In-Context Lexical Similarity. This resource proposes two subtasks in 4 different languages, English, Finnish, Croatian and Slovene. Each entry of CoSimLex is a pair of target words, and two different contexts in which both words appear. For each context is provided a measure of lexical similarity between the two words. An example of a CoSimLex entry in English is provided in Table 3, where the respective meanings of *put* and *give* are closer in context 2 than in context 1.

Tasks. Subtask 1, *Contextual Changes*, aims at predicting the change of similarity between the two words, from context 1 to context 2. This is measured with Pearson Correlation coefficient between the differences of gold similarities and the differences of predicted similarities. Subtask 2, *Similarity Ratings*, adopts a more traditional setting, where candidate systems are simply tasked to predict a measure of lexical similarity between the two words in each context, and measuring the Spearman Correlation between gold measures of similarity and predicted ones. Subtask 2 uses independently each context provided in Subtask 1 instead of pairing them, and as such covers twice as many examples.

Predicting lexical similarity in the same-context setting. For both subtasks, given a context from CoSimLex where two target words appear, we compute the embeddings corresponding to each target and then compare them using cosine distance. For XLM-R, we take the embeddings corresponding to the two sets of subwords, that compose each target. For XL-LEXEME and CALE, we create two sepa-

SUBTASK 1 : CONTEXTUAL CHANGES	EN	FI	HR	SL
# examples	340	24	112	111
XLM-R (first)	.18	.32†	.20	.33
XLM-R (mid)	.69*	.36†	.58	.52
XLM-R (last)	.63	.27†	.51	.45
XL-LEXEME	.70	.47*	.67*	.66*
CALE (XLM-R)	.74*	**.63**	**.72**	**.70**
CALE (MBERT)	.68	-	-	-
CALE (XL-LEX)	**.74**	.60*	.72*	.67*

SUBTASK 2 : SIMILARITY RATINGS	EN	FI	HR	SL
# pairs	680	48	224	222
XLM-R (first)	.27	-.04†	.25	.25
XLM-R (mid)	.61	.25†	.53	.51
XLM-R (last)	.59	.16†	.52	.46
XL-LEXEME	.63	.59*	.64	.67*
CALE (XLM-R)	**.67**	**.64**	.68*	**.68**
CALE (MBERT)	.67*	-	-	-
CALE (XL-LEX)	.67*	.57*	**.73**	.68*

Table 4: Pearson Correlation of CoSimLex Subtask 1 (upper table) and Spearman Correlation of CoSimLex Subtask 2 (lower table). Values marked with † are non-significant correlation values. The highest score for each dataset is in bold, and other scores that are not significantly different from it are marked with *.

	Silh (S)	Silh (UB)	WuP ρ
XLM-R (first)	-0.166	-0.267	0.187
XLM-R (mid)	-0.064	-0.261	0.338
XLM-R (last)	-0.088	-0.260	0.311
XL-LEXEME	-0.124	-0.370	0.489
CALE (XLM-R)	-0.041	-0.192	0.509
CALE (MBERT)	**0.042**	**-0.104**	0.514
CALE (XL-LEX)	0.003	-0.228	**0.532**

Table 5: Silhouette scores with synsets (S) or with Unique Beginners (UB) as cluster labels ; and Spearman Correlation of cosine similarities with Wu-Palmer's similarities in WordNet.

rate input sequences: the first input contains the full context with our delimiter tags placed around the first target word only; and the second with delimiters around the second target only. Feeding each sequence to the model, we obtain two embeddings, each representing the meaning of the specifically tagged target word.

Results and discussion. Results are displayed in Table 4. CALE models reach the highest correlations values in all languages, ranging between 0.63 to 0.73 in subtask 1 and 0.57 to 0.68 in subtask 2, and consistently improve over XLM-R. Surprisingly, XL-LEXEME also performs very well and is most often not deemed significantly different from CALE results, despite being trained only on a same-lemma task and not on inter-words relations. CALE (XL-LEX) does not perform significantly better than the more simple CALE (XLM-R), indicating that CALE fine-tuning is usually enough to obtain accurate in-context similarities. Again, the good performances reached by CALE fine-tuning highlight the cross-lingual transferability of XLM-R models.

6 Spatial Organization

We investigate how CALE fine-tuning affects representation space structure. Figure 1 shows cosine distance distributions across the *test* split of SPCD by pair category, while Table 5 analyzes how well embedding spaces reflect WordNet's conceptual organization using synsets annotations.

Distance distributions. In Figure 1 we plot the distributions of distance in pairs according to their categories: 'SL' and 'DL' refer to "same lemma" and "different lemmas" respectively, 'SC' and 'DC' to "same concept" and "different concepts" respectively. We first observe that XLM-R suffers from major anisotropy (i.e. all cosine distances are low

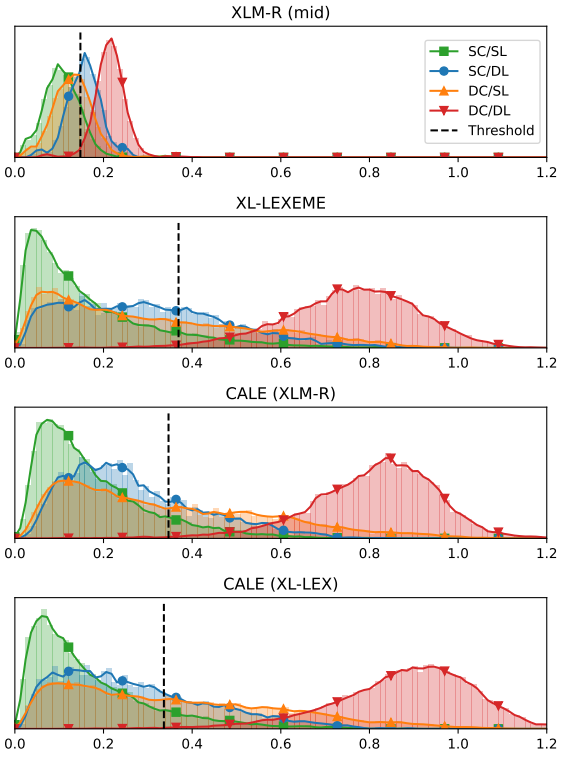

Figure 1: Distribution of cosine distance between embeddings from the candidate models' spaces for SPCD *test* pairs, according to their category. The dashed line is the decision threshold for Concept Differentiation learned from the other splits.

even in unrelated pairs of the DC/DL category). The fine-tuning of CALE and XL-LEXEME pushes occurrences in DC/DL pairs away from one another, a consequence of the margin parameter. Another observation is that CALE's fine-tuning successfully shifts the focus from *lemma*-centric to *concept*-centric: in `XLM-R`, DC/SL pairs (cases of polysemy) are generally closer than SC/DL pairs (cases of synonymy); this tendency is reversed in `CALE (XLM-R)` (average distance of .130 for DC/SL vs .157 for SC/DL before fine-tuning; .343 vs .269 after fine-tuning). Yet, we remark that the cosine distances for several DC/SL pairs (examples of the polysemy of a single lemma) remain low: distances in the DC/SL category are not distributed in a tight mass, but rather a large band across the cosine distance range. We hypothesize that the reason for this low cosine distance value is that this corresponds to pairs of similar senses. Senses of a given lemma can be very similar to one another, and WordNet is known to be very fine-grained in its sense distinctions. Indeed, after examining this phenomenon more closely, we have found a (significant) correlation coefficient of 0.2 between Wu-Palmer similarity in WordNet and cosine distances between

CALE (XL-LEX) embeddings. Thus, surprisingly low cosine distances in DC/SL pairs may be explained by similarity between word senses in the reference resource directly (at least partially, given that the correlation value is significant but not very high).

Similarity correlations. WordNet's synsets are linked to one another through relations of hyper-/hyponymy, in a tree-like structure. A short path between two synsets indicate that the two concepts are similar. We compute Spearman correlations between model's cosine similarities and Wu-Palmer similarities (Wu and Palmer, 1994) between synsets for each occurrence pair[5]. CALE models achieve the highest correlations (above 0.5), indicating a strong alignment between embedding similarity and semantic similarity. All correlations and pairwise differences are found statistically significant.

Cluster analysis. We evaluate whether occurrences of the same concept form tight, well-separated clusters using the Silhouette score (Rousseeuw, 1987), with WordNet synsets as labels. Silhouette score measures how easily occurrences would be assigned to their cluster by comparing the distance to the occurrence's cluster to the distance to the closest other cluster. A score close to 1 indicates that cluster are well-formed and well-separated. Scores close to 0 indicate that there exist overlaps between clusters, and scores close to -1 indicate that occurrences are almost systematically assigned to the wrong cluster. CALE models yield the highest scores, suggesting better conceptual clustering. However, all scores remain below 0.1, highlighting substantial overlap between clusters even for the best models, which is likely due to the fine granularity of WordNet synsets.

Broader clusters. We repeat the clustering analysis using WordNet's Unique Beginners (UB) broad semantic categories like "entity", "feeling", or "communication" as labels, limited to nouns and verbs. Overall scores are negative, indicating poor alignment with these higher-level groupings. Nonetheless, CALE performs best among models, suggesting a modest ability to reflect coarse semantic distinctions.

Conclusions. These findings show that CALE embeddings align more closely with WordNet's structure than other models, both in distances and clustering. Notably, ModernBERT-based CALE shows stronger structural alignment than multilingual models. XL-LEXEME, despite good pairwise similarity, shows weaker clustering, presumably because its fine-tuning was sense-level and did not require broad conceptual organization.

[5]We also tried Lin similarity, with similar results.

7 Conclusion

We introduced Concept Differentiation, a supervised task for predicting whether two occurrences share the same semantic concept. We curated the SPCD dataset from SemCor and developed CALE models using contrastive fine-tuning of SentenceBERT-style architectures.

Evaluation shows that CALE improves lexical semantic representations across tasks, including Concept Differentiation, Lexical Semantic Change Detection, and Lexical Similarity. These gains hold for both intra-lemma and inter-lemma tasks, with CALE matching or surpassing strong baselines like XL-LEXEME, even on multilingual benchmarks despite English-only training. This suggests that Concept Differentiation subsumes other lexical tasks dealing with the mapping between words and meanings, and that a broader view is beneficial for multi-purpose semantic representations.

Analysis of CALE's embedding geometry reveals a shift from lemma-centric to concept-centric representations and closer alignment with WordNet's organization. These results position Concept Differentiation as a promising general-purpose training signal for lexical semantics.

Future work could explore specifically-curated datasets for Concept Differentiation, with human annotators and a different annotation scheme than WordNet's synsets. We also look forward to applications of CALE to Multiword Expressions (MWE), and potential applications to Frame Induction, Concept Induction or other semantic tasks.

Limitations

While CALE models demonstrate cross-lingual transferability, fine-tuning was performed exclusively on English language data. Even if multilingual models (e.g., XLM-R) generalize well to other languages in tasks like LSCD and CoSimLex, the performance in lower-resource or typologically distant languages has not been thoroughly evaluated due to the lack of data sources. A more rigorous multilingual training and evaluation setup is needed to assess CALE's robustness across languages.

Second, the binary nature of Concept Differenti-

ation may oversimplify semantic similarity, which is often graded rather than categorical. This could limit the model's applicability in tasks requiring nuanced semantic judgments.

Third, while we show improved alignment with WordNet's conceptual structure, our models receive no direct supervision about ontological hierarchy or fine-grained relations between concepts (e.g., hyper-/hyponymy, holo-/meronymy, antonymy). Further work is needed to evaluate and potentially incorporate structured knowledge explicitly during training.

Lastly, we relied on small architectural modifications compared to XL-LEXEME for model comparison, a key aspect of our studies is model and data scaling. However, other scaling axes, notably in terms of model parameters and embedding dimension size are left unexplored.

Acknowledgments

We gratefully thank the anonymous reviewers for their insightful comments. This research was funded by Inria Exploratory Action COMANCHE. We would like to thank Pascal Denis and Mikaela Keller for their valuable feedback on the early draft of this paper.

References

Marianna Apidianaki and Aina Garí Soler. 2021. ALL dolphins are intelligent and SOME are friendly: Probing BERT for nouns' semantic properties and their prototypicality. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 79–94, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France. European Language Resources Association.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4330–4338. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Max Boholm, Björn Rönnerstrand, Ellen Breitholtz, Robin Cooper, Elina Lindgren, Gregor Rettenegger, and Asad Sayeed. 2024. Can political dogwhistles be predicted by distributional methods for analysis of lexical semantic change? In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 144–157, Bangkok, Thailand. Association for Computational Linguistics.

Pierluigi Cassotti, Pierpaolo Basile, and Nina Tahmasebi. 2024. DWUGs-IT: Extending and standardizing lexical semantic change detection for Italian. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 190–197, Pisa, Italy. CEUR Workshop Proceedings.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 227–244, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Honolulu, Hawaii. Association for Computational Linguistics.

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

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aina Garí Soler and Marianna Apidianaki. 2020. MULTISEM at SemEval-2020 task 3: Fine-tuning BERT for lexical meaning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 158–165, Barcelona (online). International Committee for Computational Linguistics.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2023. Measuring lexico-semantic alignment in debates with contextualized word representations. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 50–63, Toronto, Canada. Association for Computational Linguistics.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Martin Haspelmath. 2023. Coexpression and synexpression patterns across languages: comparative concepts and possible explanations. *Frontiers in Psychology*, 14.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.

David Jurgens and Ioannis Klapaftis. 2013. SemEval-2013 task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, Georgia, USA. Association for Computational Linguistics.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden. Association for Computational Linguistics.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Yuxi Li, Emmanuele Chersoni, and Yu-Yin Hsu. 2024. Investigating aspect features in contextualized embeddings with semantic scales and distributional similarity. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 80–92, Mexico City, Mexico. Association for Computational Linguistics.

Bastien Liétard, Pascal Denis, and Mikaela Keller. 2024. To word senses and beyond: Inducing concepts with contextualized language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2684–2696, Miami, Florida, USA. 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.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 task 14: Word sense induction &disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Anna Mosolova, Marie Candito, and Carlos Ramisch. 2024. Injecting Wiktionary to improve token-level contextual representations using contrastive learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 34–41, St. Julian's, Malta. Association for Computational Linguistics.

Gregory L Murphy. 2002. *The big book of concepts*. MIT Press.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Behrang QasemiZadeh, Miriam R. L. Petruck, Regina Stodden, Laura Kallmeyer, and Marie Candito. 2019. SemEval-2019 task 2: Unsupervised lexical frame induction. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 16–30, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

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.

James H Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2021. Semantic frame induction using masked word embeddings and two-step clustering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 811–816, Online. Association for Computational Linguistics.

Lei Yu and Yang Xu. 2023. Word sense extension. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3281–3294, Toronto, Canada. Association for Computational Linguistics.

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.

Zhuosheng Zhang, Shuohang Wang, Yichong Xu, Yuwei Fang, Wenhao Yu, Yang Liu, Hai Zhao, Chenguang Zhu, and Michael Zeng. 2022. Task compass: Scaling multi-task pre-training with task prefix. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5671–5685, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

A SPCD Dataset

To extract occurrences from the SemCor corpus Fellbaum (1998), we proceeded as follows:

- We used SemCor as distributed through the NLTK[6] library.
- We only used SemCor sentences that had semantic annotations (of WordNet's synsets) and contained minimum 10 words and no more than 100.
- For each sentence, we extract an occurrence for each tagged word in the sentence (retrieved occurrences are delimited by the sentence).
- We focus on Adjective, Nouns and Verbs, using the Part-of-Speech tag indicated by the WordNet's synsets. We merge PoS-tags "a" and "s" as they both cover Adjectives.
- For each PoS independently (to filter-out potential noisy PoS annotations), we keep only lemmas that meet the following conditions:
  - has minimum 3 letters.
  - occurs at least 10 times with the PoS.
  - is a single word (no compound) and contain only letters (no symbol/number).
  - is not a proper noun.
- We discard all occurrences of non-selected lemmas for each PoS tag, and then join the 3 remaining sets of lemmas and their occurrences. Now that extraction and filtering are done, we no longer differentiate on PoS tags (we allow pairs of occurrences from different PoS tags).

Table 8 provides the number of pairs in each split and category.

B Fine-Tuning Hyperparameters

We conducted hyperparameter optimization using Optuna's[7] Tree-structured Parzen Estimator (TPE) (Bergstra et al., 2011) algorithm to identify the best-performing configuration for our model. The optimal hyperparameters found through this search are presented in Table 6.

Hyper-parameter	Value
Margin	0.7
Learning Rate	6.02e-06
Warmup Ratio	0.24
Weight Decay	0.05
Adam β_1	0.9
Adam β_2	0.999
Adam ϵ	1e-08
Epochs	1
Seed	42
Embedding Size	1024

Table 6: CALE models hyper-parameters.

C Test Metrics in SPCD

Table 9 presents more detailed evaluation results on the SPCD dataset, distinguishing between the different pair categories (Same concept/Different concepts and Same lemma/Different lemma).

Table 7 shows the F1 score in the *test* split of SPCD.

D APD and PRT results on DWUGs

In addition to APD, we also tried another measure from Kutuzov and Giulianelli (2020), the Prototype Distance (PRT), defined as follows:

$$\mathrm{PRT}(w) = d\left(\frac{1}{n}\sum_{i=1}^{n} e^1_{w,i}\ ,\ \frac{1}{m}\sum_{j=1}^{m} e^2_{w,j}\right)$$

with d the cosine distance operator, $e^k_{w,i}$ the embedding of the i-th occurrence of word w at time T_k,

[6] https://www.nltk.org/install.html

[7] https://optuna.org/

	All	SL	DL
1L1C (baseline)	62.8	69.5	0.0
XLM-R (first)	50.2	57.8	0.0
XLM-R (mid)	67.1	69.7	53.3
XLM-R (last)	64.2	68.2	42.8
XL-LEXEME	73.5	72.5	78.2
CALE (XLM-R)	**76.2**	74.5	**83.6**
CALE (MBERT)	75.4	74.4	79.9
CALE (XL-LEX)	76.0	**74.8**	81.2

Table 7: F1 scores on test pairs from SemCor. "SL" (resp. "DL") refers to pairs of occurrences of the same lemma (resp. of different lemmas) and "SC" (resp. "DC") to pairs of occurrences referring to the same concept (resp. to different concepts).

for a target word w with n occurrences at T_1 and m occurrences at T_2.

Side-by-side results of APD and PRT can be found in Table 10.

E Hardware and Code

We conducted all experiments with Nvidia A30 GPU card with 24GB memory and Intel Xeon Gold 5320 CPU. The main libraries used include Pytorch 2.5.1, HuggingFace transformers 4.48.1, datasets 3.2.0 and sentence-transformers 3.3.1. Total training time for CALE models ranges from 16-20 hours including hyper-parameter search. Evaluation time ranges approximately from 1-2 hours.

F Scientific Artifacts

We used WordNet and SemCor, both properties of Princeton University. Licence can be found at https://wordnet.princeton.edu/license-and-commercial-use.

For DWUG datasets:

- Italian and Latin datasets are licenced under CC-BY 4.0 (https://creativecommons.org/licenses/by/4.0/).
- English, German, Spanish, Swedish datasets are licenced under CC-BY-ND 4.0 (https://creativecommons.org/licenses/by-nd/4.0/).

The CoSimLex dataset, distributed at https://huggingface.co/datasets/cjvt/cosimlex, is published under the GNU GPL 3.0 licence https://www.gnu.org/licenses/gpl-3.0.html.

The SPCD dataset was derivated from the material of SemCor and WordNet and is intended for research use.

	Total	Concept		Lemma		Cross-categories			
Split		**Same**	**Different**	**Same**	**Different**	**SC&SL**	**SC&DL**	**DC&SL**	**DC&DL**
Train	156,387	63,752 *(41%)*	92,635 *(59%)*	91,551	64,836	48,517	15,235	43,034	49,601
Val	20,346	8,114 *(40%)*	12,232 *(60%)*	12,115	8,231	6,387	1,727	5,728	6,504
Test	44,891	18,317 *(41%)*	26,574 *(59%)*	26,318	18,573	14,018	4,299	12,300	14,274

Table 8: SPCD dataset description in number of pairs. The train split used 49,601 unique occurrences, 6,504 for validation and 14,274 for test."SL" (resp. "DL") refers to pairs of occurrences of the same lemma (resp. of different lemmas) and "SC" (resp. "DC") to pairs of occurrences referring to the same concept (resp. to different concepts).

	All†	**SL**†	**DL**†	**SC&SL**	**DC&SL**	**DC&DL**	**SC&DL**
1L1C (baseline)	65.1	50.0	50.0	100	0.0	100	0.0
XLM-R (first)	60.8	54.2	50.0	54.1	51.5	100	0.0
XLM-R (mid)	70.9	59.1	68.4	84.6	30.4	95.5	39.5
XLM-R (last)	69.1	59.6	63.5	77.9	38.4	95.8	28.8
XL-LEXEME	76.7	62.4	82.7	89.7	35.0	98.2	67.4
CALE (XLM-R)	79.3	65.9	86.4	89.1	38.0	98.5	73.1
CALE (MBERT)	78.7	66.6	83.5	87.2	41.9	98.9	67.0
CALE (XL-LEX)	79.2	67.2	84.4	87.2	43.8	98.9	69.4

Table 9: Accuracies on test pairs from SemCor (Balanced Accuracies are indicated with †). The higher the better. "All" refers to all pairs, "SL" (resp. "DL") to pairs of occurrences of the same lemma (resp. of different lemmas) and "SC" (resp. "DC") to pairs of occurrences referring to the same concept (resp. to different concepts).

	EN		**DE**		**ES**		**IT**		**LA**		**SV**	
	APD	PRT	APD	PRT	APD	PRT	APD	PRT	APD	PRT	APD	PRT
XLM-R (first)	.221	.313	.143	.307	.078	.308	.180	.327	.004	.196	.098	.031
XLM-R (mid)	.568	.500	.543	.756	.495	.510	.316	.538	**.237**	.481	.575	.466
XLM-R (last)	.564	.527	.342	.404	.526	.562	.270	.605	.001	.521	.518	.307
XL-LEXEME	**.786**	**.671**	.798	.803	.633	.610	.512	.637	.142	**.549**	.854	.739
CALE (XLM-R)	.705	.533	.777	**.823**	.655	.543	**.755**	**.753**	.108	.461	.833	.720
CALE (MBERT)	.655	.527	-	-	-	-	-	-	-	-	-	-
CALE (XL-LEX)	.730	.570	**.826**	.799	**.685**	**.639**	.543	.676	.195	.485	**.859**	**.766**

Table 10: Spearman correlations between gold change scores and predicted change scores for LSCD on the DWUG datasets, for APD and PRT measures.