GrEmLIn: A Repository of Green Baseline Embeddings for 87 Low-Resource Languages Injected with Multilingual Graph Knowledge

Daniil Gurgurov[1] **Rishu Kumar**[1] **Simon Ostermann**[1,2]
[1]German Research Center for Artificial Intelligence (DFKI)
[2]Centre for European Research in Trusted AI (CERTAIN)
{daniil.gurgurov, rishu.kumar, simon.ostermann}@dfki.de

Abstract

Contextualized embeddings based on large language models (LLMs) are available for various languages, but their coverage is often limited for lower resourced languages. Using LLMs for such languages is often difficult due to a high computational cost; not only during training, but also during inference. Static word embeddings are much more resource-efficient ("green"), and thus still provide value, particularly for very low-resource languages. There is, however, a notable lack of comprehensive repositories with such embeddings for diverse languages. To address this gap, we present GrEmLIn, a centralized repository of green, static baseline embeddings for 87 mid- and low-resource languages. We compute GrEmLIn embeddings with a novel method that enhances GloVe embeddings by integrating multilingual graph knowledge, which makes our static embeddings **competitive with LLM representations, while being parameter-free at inference time**. Our experiments demonstrate that GrEmLIn embeddings outperform state-of-the-art contextualized embeddings from E5 on the task of lexical similarity. They remain competitive in extrinsic evaluation tasks like sentiment analysis and natural language inference, with average performance gaps of just 5-10% or less compared to state-of-the-art models, given a sufficient vocabulary overlap with the target task, and underperform only on topic classification. Our code and embeddings are publicly available at https://github.com/d-gurgurov/GrEmLIn-Green-Embeddings-LRLs[1].

1 Introduction

Word embedding methods have revolutionized natural language processing (NLP) by capturing semantic relationships between words using co-occurrence statistics from large text corpora (Mikolov et al., 2013a; Pennington et al., 2014; Bojanowski et al., 2017). This data-driven approach has significantly improved performance across numerous NLP tasks (Lample et al., 2017; Xie et al., 2018; Almeida and Xexéo, 2019).

While contextual representations like the ones based on BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and GPT (Radford et al., 2019) nowadays provide better performance than static embeddings in many tasks, their training is computationally expensive (Strubell et al., 2019; Bommasani et al., 2021) and ineffective for data-scarce languages due to their data hunger and the curse of multilinguality (Conneau et al., 2020). Some approaches of efficient adaptation of especially large language models (LLMs) to languages other than English have been investigated in recent years (Pfeiffer et al., 2020; Vykopal et al., 2024). However, even such approaches still require hardware during runtime, as embeddings need to be computed based on a forward pass for each new text that is processed. This is often prohibitive in low-resource (hardware) scenarios, and inefficient in terms of energy use. Also, such approaches are often not tailored to low-resource languages.

In contrast, static word embeddings continue to play a crucial role in specific tasks such as bias detection and removal (Gonen and Goldberg, 2019; Manzini et al., 2019), explaining word vector spaces (Vulić et al., 2020b; Bommasani et al., 2020), and information retrieval (Yan et al., 2018). Static word embeddings have the advantage of being **parameter-free** at inference time, as no neural network needs to be loaded for computing such representations; just a dictionary lookup is required. This makes them both attractive for low-resource hardware scenarios, and **much more environment-friendly** (Strubell et al., 2019; Dufter et al., 2021). Existing resources for multilingual embedding data bases (Ferreira et al., 2016;

[1]All vectors are available on Huggingface as single model pages. Each page starts with *DFKI/glove*.

Bojanowski et al., 2017; Grave et al., 2018a) often suffer from limited scope and outdated data, potentially worsening their ability to capture the dynamic nature of language and adequately support low-resource languages. We want to fill this gap by providing GrEmLIn, a large database of static word embeddings for 87 mid- and low-resource languages.

As for LLMs, the training of word embeddings suffers from the lack of high-quality data in low-resource languages (to a smaller degree). Incorporating other types of data for improving word representations is thus beneficial especially for low-resource languages. Knowledge graphs provide such an alternative to textual knowledge, with rich semantic and multilingual sources of information, including synonyms, antonyms, morphological forms, definitions, etimological relations, translations, and more (Miller, 1995; Navigli and Ponzetto, 2012; Speer et al., 2017). Such structured and cross-lingual information can be used to improve the quality of classical word representations (Faruqui et al., 2014; Sakketou and Ampazis, 2020), which are only trained on co-occurence statistics.

To that end, we propose a new simple yet effective method for including graph information into word embeddings based on Mikolov et al. (2013b). We learn a projection matrix to map static embeddings to a combined space, effectively overcoming the limitations of retrofitting approaches that only enhance a limited vocabulary. This method combines the strengths of traditional word embeddings with the structured, multilingual information from knowledge graphs, resulting in more accurate and informative representations.

In summary, our contributions in this work are two-fold: First, we present GrEmLIn, a centralized resource of static word embeddings for 87 mid- and low-resource languages, specifically focusing on word embeddings trained with GloVe (Pennington et al., 2014). Second, we propose an effective method to improve embeddings by incorporating more knowledge in the form of multilingual knowledge graphs, which is especially important for low-resource languages, where resources are usually very scarce. Our code is publicly available on GitHub[2].

[2] https://github.com/d-gurgurov/GrEmLIn-Green-Embeddings-LRLs

2 Related Work

We briefly describe the most prominent graph knowledge sources, word embeddings, and existing methods for improving embeddings with graphs.

Graph knowledge sources. Among most used knowledge graphs for natural language are WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2012). WordNet is a lexical database that organizes English words into sets of synonyms called synsets, providing short definitions and usage examples. BabelNet is a multilingual encyclopedic dictionary and semantic network, which integrates lexicographic and encyclopedic knowledge from WordNet, Wikipedia, etc., focused on named entities. In our work, we use ConceptNet (Speer et al., 2017), a multilingual, domain-general knowledge graph that connects words and phrases from various natural languages with labeled, weighted edges representing relationships between terms. Unlike other knowledge graphs, ConceptNet is not a monolingual collection of named entities but focuses on commonly used words and phrases across multiple languages.

Word embeddings. *Word2Vec* (Mikolov et al., 2013a) uses shallow neural networks to produce word vectors. It comes in two types: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts a word given its context, while Skip-gram predicts the context given a word. *GloVe* (Global Vectors for Word Representation) (Pennington et al., 2014) word embeddings are created by aggregating global word-word co-occurrence statistics from a corpus. The resulting vectors capture both local and global semantic relationships. *FastText* (Bojanowski et al., 2017) extends Word2Vec by representing words as bags of character n-grams, capturing subword information and handling out-of-vocabulary words more effectively. FastText is particularly useful for morphologically rich languages. *Numberbatch*, part of the ConceptNet project (Speer et al., 2017), is a set of word embeddings that integrates knowledge from ConceptNet with distributional semantics from GloVe and Word2Vec. Numberbatch uses a retrofitting approach (Faruqui et al., 2014) to enhance embeddings with structured semantic knowledge. Retrofitting often results in a limited vocabulary for underrepresented languages (Speer and Lowry-Duda, 2017) since the retrofitting process relies on existing semantic relationships in ConceptNet to adjust the original embeddings.

Improving Embeddings with Knowledge Graphs. There are various methods to improve word embeddings by incorporating external knowledge graphs or semantic networks (Dieudonat et al., 2020). *Retrofitting* (Faruqui et al., 2014) is a post-processing technique that adjusts pre-trained word embeddings using information from knowledge graphs or semantic lexicons. The key idea is to infer new vectors that are close to their original embeddings while also being close to their neighbors in the graph or lexicon. This is achieved by minimizing an objective function that balances the distance between the new vectors and the original embeddings, as well as the distance between connected nodes. *Expanded retrofitting* (Speer et al., 2017), used for ConceptNet Numberbatch, optimizes over a larger vocabulary including terms from the knowledge graph not present in the original embeddings, but it still does not retrofit all the words in the original embedding space. Other existing methods that integrate contextualized embeddings with knowledge graph embeddings often use attention mechanisms, as demonstrated by works such as Peters et al. (2019), Zhang et al. (2019), and Gurgurov et al. (2024). These methods specifically enhance BERT embeddings by incorporating external knowledge bases.

3 Method

We propose a method for merging GloVe embeddings with graph-based embeddings derived from ConceptNet knowledge, while preserving the vocabulary size of GloVe, following two steps: First, we use singular value decomposition (SVD) (Eckart and Young, 1936) on concatenated word embeddings from GloVe and pointwise mutual information (PMI) based graph embeddings (Speer et al., 2017) to generate a shared embedding space. We do so for the part of the vocabulary that is shared between GloVe and the knowledge graph. Second, we learn a linear transformation from GloVe into this joined space to obtain embeddings for all words in the original GloVe vocabulary.

3.1 GloVe Embeddings

We train GloVe embeddings using the original code. The model is trained by stochastically sampling nonzero elements from the co-occurrence matrix over 100 iterations, to produce 300-dimensional vectors. We use a context window of 10 words to the left and 10 words to the right. Words with fewer than 5 co-occurrences are excluded for languages with over 1 million tokens in the training data, and the threshold is set to 2 for languages with smaller datasets. We use data from CC100[3] (Wenzek et al., 2020; Conneau et al., 2020) for training the static word embeddings. We set $x_{max} = 100$, $\alpha = \frac{3}{4}$, and use AdaGrad optimization (Duchi et al., 2011) with an initial learning rate of 0.05.

3.2 Graph Embeddings

To build ConceptNet-based word embeddings, we follow the method used for constructing ConceptNet Numberbatch embeddings (Speer et al., 2017). We represent the ConceptNet graph as a sparse, symmetric term-term matrix, where each cell is the sum of the occurences of all edges connecting the two terms. Unlike the original method, we do not discard terms connected to fewer than three edges, as we deal with low-resource languages.

We calculate embeddings from this matrix by applying pointwise mutual information (PMI) with context distributional smoothing of 0.75, clipping negative values to yield positive PMI (PPMI), which follows practical recommendations by (Levy et al., 2015). We then reduce the dimensionality to 300 using truncated SVD and combine terms and contexts symmetrically to form a single matrix of word embeddings, called ConceptNet-PPMI. This matrix captures the overall graph structure of ConceptNet.

We compute ConceptNet-PPMI embeddings for the entire ConceptNet, covering 304 languages, which we call *PPMI (All)*. Further, we construct separate graph embedding spaces, *PPMI (Single)*, for each language, using only the portion of ConceptNet for that language. This approach is adopted since the initial co-occurence matrices for individual languages are less sparse while still being multilingual in nature.

3.3 Singular Value Decomposition (SVD)

We first concatenate GloVe and PPMI vectors for all words that are in the shared vocabulary, resulting in 600-dimensional vectors[4]. Afterwards, we reduce the dimensionality and remove some

[3] https://huggingface.co/datasets/cc100

[4] PPMI embeddings are normalized to be in the range of the Glove embeddings

of the variance coming from redundant features. The matrix M representing merged GloVe and ConceptNet-PPMI can be approximated with a truncated SVD:

$$M \approx U \Sigma V^T$$

where Σ is truncated to a $k' \times k'$ diagonal matrix of the k' largest singular values, and U and V are correspondingly truncated to have only these k' columns. U is then used as a matrix mapping the original vocabulary to a smaller set of features[5].

3.4 Linear Transformation

To obtain embeddings for the entire vocabulary from the original GloVe embedding space (i.e. not only the common words), we find a linear projection matrix between the spaces and project the GloVe embeddings onto the merged embedding space, similar to Mikolov et al. (2013c), using a gradient descent optimization on a linear regression model.

Given a set of word pairs and their associated vector representations $\{x_i, z_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^{d_1}$ is the GloVe representation of word i, and $z_i \in \mathbb{R}^{d_2}$ is the PPMI representation from ConceptNet, our goal is to find a transformation matrix W such that Wx_i approximates z_i.

W can be learned by solving the following optimization problem:

$$\min_W \sum_{i=1}^{n} \|Wx_i - z_i\|^2$$

which we solve as a linear regression problem with stochastic gradient descent optimization.

The resulting projection matrix is used to project the GloVe embeddings onto the merged embedding space.

4 Experiments

In this section, we describe the selected languages, tasks, and experiments conducted to evaluate the effectiveness of our proposed method.

4.1 Languages

We trained GloVe embeddings for 87 languages from the CC100 dataset (Wenzek et al., 2020), focusing on languages categorized as low-resource (class 0 to 3) based on Joshi et al.'s classification (2020). For 72 of these languages, present in both CC100 and ConceptNet, we generated additional graph embeddings. The merging process involves enhancing the original GloVe embeddings with graph knowledge via SVD-reduced PPMI integration. Further details about these languages, including common vocabulary size between GloVe and ConceptNet, can be found in Part C of the Appendix.

4.2 Evaluation Data

We assess the embeddings using both intrinsic and extrinsic evaluation tasks. The intrinsic evaluation is performed using the MultiSimLex dataset (Vulić et al., 2020a), which provides manually annotated data on semantic similarity consisting of 1888 examples across 12 languages, 4 of which overlap with our work. This task focuses on measuring the strength of similarity between word pairs (e.g., "lion – cat") independently of relatedness, making it a good test for how well embeddings capture semantic similarity.

For extrinsic evaluations, we focus on three downstream NLP tasks: Sentiment Analysis (SA), Topic Classification (TC), and Natural Language Inference (NLI). Due to the limited availability of intrinsic datasets for most low-resource languages, we prioritize these tasks to reflect real-world use cases, where high-quality word embeddings are crucial.

For SA, we compile data for 23 languages from multiple open sources, prioritizing mid- and low-resource languages for broader coverage across typological families. The details of these data sources are listed in Table 7 in the Appendix. Some datasets, such as those for Swahili, Nepali, Uyghur, Latvian, Slovak, Slovenian, Uzbek, Bulgarian, Yoruba, Bengali, Hebrew, and Telugu, are highly imbalanced in terms of class distribution. To mitigate this, we apply random undersampling to create a balanced version of the datasets. This step allows for a more robust comparison of the embeddings' performance in low-data settings.

We further evaluate the embeddings on the TC task using the SIB-200 dataset (Adelani et al., 2024). It offers multilingual data for topic classification, covering 200 languages, and was specifically designed to improve natural language understanding for under-resourced languages. Our experiments cover 57 languages, chosen based on their availability in both ConceptNet and CC100.

[5]We dismiss the weighting of U by the singular values from Σ, which was noted to work better for semantic tasks (Levy et al., 2015)

The task is framed as a multi-label classification with the data distributed along 7 different classes. The dataset provides predefined train, validation and test splits, which consist of 701, 99, and 204 examples, respectively.

Lastly, we evaluate the embeddings on the NLI task, using the XNLI dataset (Conneau et al., 2018). The XNLI dataset provides multilingual NLI examples for 15 languages, and for our experiments, we selected 5 of these languages: Swahili, Urdu, Greek, Thai, and Bulgarian. These languages were selected based on the availability of our GloVe, PPMI-enhanced embeddings, and the NLI dataset. Evaluating embeddings on the NLI task tests their ability to understand logical relationships between sentence pairs, an important capability for higher-level NLU tasks. Due to the simplicity of our models, we only utilize validation and test splits, consisting of 2,490 and 5,010 examples, respectively, for training and testing, excluding the original training split of nearly 400,000 examples.

4.3 Experimental Setup

We evaluate the embeddings using a Support Vector Machine (SVM) classifier (Boser et al., 1992) for all extrinsic tasks—SA, TC, and NLI—reporting macro-averaged F1 scores for fair comparison. For the intrinsic MultiSimLex task, we use Spearman's Rank Correlation (Spearman, 1961) to assess how well the embeddings' similarity predictions align with human annotations.

For extrinsic tasks, sentence representations are constructed by summing word embeddings, which is a standard approach in NLP (Mikolov et al., 2013d; Bowman et al., 2015; Williams et al., 2018), and then used as input features for the SVM. The SVM model is trained with a Radial Basis Function (RBF) kernel, which is commonly used for nonlinear classification problems. The regularization parameter C is fixed at 1 for GloVe-based embeddings, balancing the trade-off between maximizing the margin and minimizing classification errors. This setup minimizes the impact of hyperparameters on the resulting scores.

For the NLI task, sentence representation follows the same method as above, but with an added step. We concatenate the sentence embeddings of the two input sentences (premise and hypothesis) to form the final input representation for the SVM. This approach enables the model to capture the relationship between the two sentences.

As for baselines, we use three strong pre-trained models:

- FastText (Grave et al., 2018b), a word embedding model that extends the traditional skip-gram model by representing words as bags of character n-grams, allowing it to effectively handle out-of-vocabulary words.

- XLM-R-base (Conneau et al., 2020), a transformer-based multilingual model. We obtain sentence embeddings by summing the model's last hidden states.

- E-5-base (Wang et al., 2024), a state-of-the-art multilingual sentence embedding model known for its strong performance in multilingual tasks. This serves as a high-quality benchmark for our comparisons.

For the XLM-R-base and E-5 embeddings, we adjust the regularization parameter C to 100. This adjustment accounts for the higher dimensionality of these embeddings, as lower C values constrain their performance.

5 Results

We distinguish between static and contextualized embeddings by first comparing the static embeddings against each other, and then comparing them to the contextualized ones. The results from E-5-B are provided for reference but cannot be directly compared to our static embeddings due to the reasons outlined in Section 6.

5.1 Semantic Similarity

We evaluate the performance of the embeddings on the lexical semantics task using the MultiSimLex dataset, focusing on 4 languages: Estonian, Welsh, Swahili, and Hebrew.

As shown in Table 1, the GloVe+PPMI (Single) embeddings achieve the highest correlation scores for 3 out of 4 languages, demonstrating their ability to capture semantic similarities. For Swahili, FastText achieves the best result, although GloVe+PPMI remains competitive. In contrast, contextual embeddings such as XLM-R-base struggle in this intrinsic evaluation task, achieving lower correlation scores across all languages, which supports Vulić et al. (2020a). E-5 performs better than XLM-R but does not surpass the best-performing static embeddings.

Cov.	ISO	Contextualized		Static			
		E-5-B	X-B	FT	G	GP(S)	GP(A)
>90%	et	**.19**	.03	.447	.341	**.452**	.422
	he	**.218**	.057	.426	.336	**.436**	.429
	Avg.	**.204**	.044	.437	.339	**.444**	.426
<90%	cy	**.112**	.039	.346	.276	**.366**	.357
	sw	**.212**	.011	**.408**	.24	.319	.324
	Avg.	**.162**	.003	**.377**	.258	.343	.341
	All avg.	**.183**	.034	**.407**	.298	.393	.383

Table 1: Spearman's correlation scores on MultiSimLex across 4 languages, for E-5-B, XLM-R-B, FastText, GloVe (G), GloVe + PPMI (GP), Single and All, sorted by GloVe vocabulary coverage. The horizontal solid line separates languages with over 90% coverage (above) from those with less (below). **Bold** numbers indicate the maximum per line, for static and contextualized.

These results underscore the continued relevance of graph-enhanced static embeddings in lexical semantic tasks, particularly for low-resource languages where training data may be scarce.

5.2 Sentiment Analysis

We evaluate the performance of the proposed GloVe+PPMI embeddings on the SA task for 23 mid- and low-resource languages. Table 2 presents the results for this task. Our findings show that both GloVe+PPMI (Single) and GloVe+PPMI (All) embeddings consistently outperform the original GloVe embeddings across most languages. GloVe+PPMI (Single) improves performance for 19 out of 23 languages, while GloVe+PPMI (All) improves results for 18 out of 23 languages when compared to GloVe.

When comparing GloVe with FastText embeddings, we observe that GloVe outperforms FastText in 12 out of 23 languages, with some languages showing comparable results.

In contrast, XLM-R-base performs better than all static embedding configurations for 9 out of 23 languages, and E-5 outperforms most static embedding variants. While this underscores the power of contextualized models, the enhanced GloVe+PPMI embeddings remain competitive, with a drop of only 5% in performance, especially in low-resource settings. This suggests that static embeddings, when enriched with multilingual graph knowledge, remain competitive and provide a lightweight and efficient zero-parameter alternative for resource-constrained environments.

5.3 Natural Language Inference

In the NLI task using the XNLI dataset, we again observe consistent improvements in performance with the enhanced GloVe embeddings (Table 3). While GloVe outperforms FastText for only 2 out of 5 languages, the use of PPMI (Single) and PPMI (All) results in better performance for all 5 languages.

In comparison, XLM-R performs better than the static embedding variants for 1 out of the 5 languages, and E-5 outperforms all models in all languages. While transformer models like XLM-R excel in capturing complex semantic relationships between sentences, the performance of GloVe+PPMI remains competitive, with a drop of only 6% given a sufficient vocabulary overlap, especially in improving sentence-level reasoning and inference capabilities in low-resource languages.

5.4 Topic Classification

The results of the topic classfication task using the SIB-200 dataset are the only results where the contextualized models seem to have a clear advantage. The drop for a vocabulary coverage of over 95% is only at 10%, but over all languages the drop averages at 20% when comparing the best contextualized with the best static model (Table 4). GloVe embeddings outperform FastText for 27 out of 57 languages, using GloVe+PPMI (Single) boosts performance for 37 out of 57 languages, and GloVe+PPMI (All) enhances performance for 48 out of 57 languages.

XLM-R gives better performance than all static embedding configurations for only 24 out of 57 languages, but E-5 performs better than all static embeddings, showcasing some strengths of con-

Cov.	ISO	Contextualized		Static			
		E-5-B	X-B	FT	G	GP(S)	GP(A)
>90%	ka	**.9**	.845	.855	.861	**.87**	.861
	sl	**.881**	.832	.743	.749	.779	**.788**
	ro	**.926**	.872	.803	.805	**.85**	.847
	Avg.	**.902**	.85	.8	.805	**.833**	.832
>80%	he	**.929**	.811	.782	.788	**.824**	.822
	si	**.895**	.831	.846	.848	.85	**.857**
	sw	**.773**	.665	.697	.68	.701	**.714**
	ug	**.881**	.61	.792	.746	**.811**	**.811**
	lv	**.801**	.74	.749	.783	**.787**	**.787**
	te	**.854**	.831	.798	.806	.808	**.817**
	sk	**.911**	.854	.73	.756	**.806**	.805
	mr	**.912**	.886	.888	.903	**.905**	.902
	bg	**.884**	.721	.793	.786	.801	**.805**
	mk	**.817**	.736	.682	**.716**	.711	.7
	Avg.	**.866**	.769	.776	.781	.8	**.802**
<80%	su	**.855**	.829	.805	.798	**.822**	.812
	am	**.861**	.782	.815	**.881**	.86	.88
	ne	**.666**	.519	.666	.643	.674	**.688**
	da	**.972**	.927	.895	.863	**.908**	.903
	uz	**.858**	.807	**.822**	.808	.806	.806
	bn	**.938**	.837	**.889**	.875	**.881**	.878
	ur	**.818**	.757	.678	.676	**.746**	.745
	az	**.787**	.762	**.75**	.744	.746	.745
	cy	**.834**	.795	.798	.77	.789	**.801**
	yo	**.764**	.634	.696	.721	.709	**.738**
	Avg.	**.835**	.765	.781	.778	.794	**.8**
	All avg.	**.857**	.778	.781	.783	.802	**.805**

Table 2: Macro Average F1 Scores for Sentiment Analysis per language, sorted by GloVe vocabulary coverage. Horizontal solid lines indicate 90% and 80% coverage by GloVe. **Bold** numbers indicate the maximum per line, for static and contextualized.

Cov.	ISO	Contextualized		Static			
		E-5-B	X-B	FT	G	GP(S)	GP(A)
>80%	bg	**.563**	.465	.465	.441	**.481**	.477
	el	**.546**	.455	.484	.456	**.496**	.488
	sw	**.539**	.437	.471	.438	.466	**.468**
	Avg.	**.549**	.452	.473	.445	**.481**	.478
<80%	ur	**.540**	.472	.412	.44	**.473**	.471
	th	**.538**	.461	.275	.284	.292	**.3**
	Avg.	**.539**	.467	.344	.362	.383	**.386**
	All avg.	**.545**	.458	.421	.412	**.442**	.441

Table 3: Macro Average F1 Scores for Natural Language Inference per language, sorted by GloVe vocabulary coverage. The horizontal solid line separates languages with over 80% coverage (above) from those with less (below). **Bold** numbers indicate the maximum per line, for static and contextualized.

textualized embeddings in multilingual tasks.

5.5 Additional Experiment: Graph-enhanced GloVe Improvement

To explain the improvements from injecting graph knowledge into static embeddings, we hypothesize that the size of the common vocabulary between GloVe and PPMI spaces contributes to performance gains: a larger vocabulary may lead to a better linear transformation fit, resulting in more precise projections. We investigate the relationship between common vocabulary size and embedding improvements by calculating Pearson (Cohen et al., 2009) and Spearman (Spearman, 1961) correlations across all tasks (SID-200, SA, XNLI, SimLex).

Table 5 shows the correlation coefficients for each task and embedding configuration. For SID-200, the GloVe+PPMI (Single) embeddings have a Spearman correlation of 0.364, indicating a moderate monotonic relationship between common vocabulary count and improvement. However, the Pearson correlation of 0.096 suggests a weak linear relationship. A similar pattern can be observed in the other tasks, where Spearman correlations are generally higher than Pearson, highlighting the non-linear nature of the relationship. In contrast, tasks like SimLex show high correlations in both metrics, especially in the Single configuration, with Pearson and Spearman scores of 0.879 and 0.8, respectively.

When comparing results for Single and All configurations, the Single configurations tend to show stronger correlations. The All configurations have higher vocabulary overlaps between the embedding spaces due to contributions from various languages (C of the Appendix). This is because Single configurations focus on one language, whereas All configurations include words from multiple languages, which dilutes the strength of the relationship between vocabulary overlap and performance improvement and may suggest that the Single embedding spaces provide a better representation of graph knowledge when working in a monolingual setting.

These results suggest that while improvement scores moderately depend on vocabulary coverage, the relationship isn't strictly linear. This implies that while a larger common vocabulary can enhance performance, other factors such as graph-based semantic knowledge may play a more significant role. Figure 1 in the Appendix visualizes the correlations for SA and SID-200 across all models.

6 Discussion: Contextual vs. Static Embeddings

While being far behind on the tested intrinsic task, the sentence embeddings extracted from E-5, a

Cov.	ISO	Contextualized		Static			
		E-5-B	X-B	FT	G	GP(S)	GP(A)
>95%	ro	**.891**	.707	.405	.561	.686	**.704**
	sk	**.866**	.707	.522	.67	.667	**.725**
	bg	**.869**	.751	.447	.645	.711	**.723**
	el	**.872**	.66	.387	.531	**.712**	.702
	lt	**.861**	.704	.534	.713	.775	**.797**
	uk	**.904**	.717	.542	.682	.722	**.745**
	lv	**.858**	.709	.608	.737	.732	**.742**
	sl	**.848**	.705	.544	.628	.715	**.734**
	gl	**.865**	.723	.522	.53	.663	**.699**
	da	**.864**	.724	.423	.446	**.743**	.717
	he	**.824**	.701	.67	.759	.739	**.784**
	mk	**.855**	.779	.598	.611	.694	**.719**
	ms	**.846**	.748	.634	.694	.738	**.769**
	Avg.	**.863**	.718	.526	.631	.715	**.735**
>90%	et	**.823**	.655	.583	.589	.572	**.605**
	be	**.836**	.698	**.674**	.58	.597	.621
	az	**.853**	.742	.667	.698	.668	**.711**
	eo	**.844**	.665	.57	.504	.567	**.588**
	hy	**.809**	.622	.411	.551	.609	**.644**
	kk	**.838**	.69	.64	.664	.647	**.69**
	is	**.798**	.651	.442	.423	.49	**.534**
	ka	**.763**	.694	.591	.684	**.689**	**.689**
	ur	**.804**	.648	.451	.42	.627	**.643**
	cy	**.704**	.638	.615	.564	.608	**.694**
	af	**.865**	.721	.573	.454	.56	**.59**
	si	**.809**	.682	.647	.678	.613	**.695**
	Avg.	**.812**	.676	.572	.567	.604	**.642**
>80%	tl	**.85**	.624	.656	.65	.707	**.709**
	bn	**.814**	.595	.567	.604	.617	**.681**
	ga	**.704**	.441	**.585**	.411	.532	.547
	mr	**.838**	.64	.567	.627	.608	**.676**
	ky	**.783**	.665	**.633**	.593	.571	.59
	gu	**.79**	.613	**.663**	.544	.589	.631
	ml	**.812**	.664	.641	**.651**	.574	.608
	pa	**.807**	.556	**.566**	.474	.521	.565
	kn	**.803**	.604	**.672**	.652	.581	.658
	ne	**.796**	.699	.553	.542	.563	**.605**
	ha	**.708**	.449	—	.421	.489	**.546**
	ja	**.802**	.608	**.656**	.511	.523	.541
	ug	**.723**	.606	**.642**	.556	.583	.622
	am	**.781**	.559	**.585**	.515	.472	.555
	Avg.	**.787**	.595	**.614**	.554	.566	.61
<80%	su	**.765**	.561	**.572**	.467	.446	.526
	so	**.642**	.388	.442	.363	.403	**.459**
	ps	**.73**	.542	**.532**	.351	.431	.493
	ht	**.717**	.318	**.531**	.392	.496	.523
	yi	**.538**	.36	**.532**	.341	.384	.453
	gd	**.54**	.341	.404	.23	.397	**.418**
	xh	**.641**	.324	—	.388	.32	.341
	yo	**.663**	.185	**.341**	.199	.211	.264
	sa	**.762**	.542	**.452**	.206	.261	.251
	qu	**.561**	.245	**.294**	.175	.167	.153
	my	**.791**	.564	.171	**.228**	.207	.163
	km	**.74**	.631	.114	**.125**	.117	.109
	ku	**.657**	.202	.09	**.11**	.098	.095
	lo	**.743**	.704	—	.183	**.185**	.18
	wo	**.594**	.238	—	.058	**.139**	.122
	Avg.	**.672**	.41	**.373**	.254	.284	.303
	All avg.	**.779**	.591	.529	.497	.535	**.566**

Table 4: Macro Average F1 Scores for Topic Classification per language, sorted by GloVe vocabulary coverage. The horizontal solid lines indicate 95%, 90%, and 80% coverage by GloVe. **Bold** numbers indicate the maximum per line, for static and contextualized.

state-of-the-art multilingual sentence embedding model, consistently outperform many other configurations across all languages on the extrinsic tasks. This superior performance can be attributed to E-5's ability to generate context-aware, sentence-level representations that capture nuanced meanings, which static embeddings, like GloVe or FastText, struggle to achieve. Unlike static word embeddings that sum individual word vectors, E-5 learns richer representations by incorporating contextual information across languages.

However, direct comparisons between E-5 and static word embeddings overlook key differences in design and use cases. E-5 is extensively trained on multilingual corpora and excels in tasks requiring complex, context-sensitive representations. In contrast, static embeddings, though simpler, are a valid alternative in low-resource or efficiency-critical scenarios, as they are effectively parameter-free during inference time. The coverage of task-specific data plays a crucial role: GloVe embeddings performe well in sentiment analysis due to broader language coverage, but poorer results in topic classification are partly linked to lower coverage in some languages. **Static embeddings remain competetive across most tested extrinsic tasks and most languages, given a good vocabulary coverage.**

Static embeddings enriched with external knowledge sources, such as graph-based information, provide significant advantages, especially in resource-limited applications where computational costs are critical. Computationally lightweight word vectors are invaluable in settings where models like E-5 are prohibitively expensive to deploy (Strubell et al., 2019; Bommasani et al., 2021). Static embeddings also perform competitively in simpler tasks that do not heavily rely on contextual understanding (Dufter et al., 2021), making them ideal for large-scale or real-time applications (Gupta and Jaggi, 2021). Additionally, static embeddings offer a level of transparency often lacking in complex models (Vulić et al., 2020b; Bommasani et al., 2020). **Their word-level semantic relationships are easy to interpret, making them useful in applications such as bias detection or model auditing.**

Furthermore, Dufter et al. (2021) demonstrated that FastText outperformed BERT on a modified LAMA task (Petroni et al., 2019) across ten languages while generating just 0.3% of BERT's car-

Task	P	S
SID-200 (Single)	0.096	0.364
SID-200 (All)	0.284	0.115
SA (Single)	0.116	0.261
SA (All)	0.254	0.186
XNLI (Single)	0.075	0.205
XNLI (All)	0.054	0.300
SimLex (Single)	0.879	0.800
SimLex (All)	0.399	0.105

Table 5: Pearson and Spearman Correlations between Common Vocabulary Count and Improvement Scores

bon footprint (Strubell et al., 2019; Dufter et al., 2021), despite their simplicity. This highlights the overlooked value of static embeddings when evaluating resource-intensive models, **rendering them useful as "green" baselines that are environmentally highly efficient**.

7 Conclusion

In this work, we developed GrEmLIn, a centralized repository of graph-enhanced GloVe embeddings for 87 mid- and low-resource languages, addressing the need for high-quality word embeddings in underrepresented languages. By merging GloVe with graph-based knowledge from ConceptNet, we enhanced the semantic richness of embeddings, leading to improved performance across tasks like semantic similarity, sentiment analysis, topic classification, and natural language inference.

Our results show that graph-enhanced GloVe outperforms the original GloVe, FastText, and even contextualized embeddings from XLM-R, offering a lightweight and environmentally efficient alternative to transformer-based models. Static embeddings have been recognized as "green" baselines, offering competitive performance at a fraction of computational cost of LLMs. This makes them ideal for low-resource settings where both computational efficiency and sustainability are key.

Limitations

While our contribution provides baseline and graph-enhanced GloVe models for many languages, several limitations exist. First, the quality and availability of training data, particularly for low-resource languages, remain key challenges. Despite leveraging large corpora like CC100 and ConceptNet, data diversity and coverage are still limited.

Second, while our method of merging GloVe embeddings with graph-based knowledge has yielded promising results, there is room for further refinement. Future work could explore more advanced fusion and projection techniques to enhance representations for low-resource languages.

Lastly, static embeddings, even with graph enhancements, cannot fully capture contextual nuances compared to transformer-based models, which may limit their performance on tasks requiring deep contextual understanding. Balancing simplicity and efficiency with improved performance remains an ongoing challenge.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. This work was supported by DisAI – Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies, a Horizon Europe-funded project under GA No. 101079164, and by the German Ministry of Education and Research (BMBF) as part of the project TRAILS (01IW24005).

References

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Felipe Almeida and Geraldo Xexéo. 2019. Word embeddings: A survey. *ArXiv*, abs/1901.09069.

Adam Amram, Anat Ben David, and Reut Tsarfaty. 2018. Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from Modern Hebrew. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2242–2252, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4758–4781, Online. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv e-prints*, pages arXiv–2108.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jože Bučar, Martin Žnidaršič, and Janez Povh. 2018. Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation*, 52(3):895–919.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Piyumal Demotte, Lahiru Senevirathe, Binod Karunanayake, Udyogi Munasinghe, and Surangika Ranathunga. 2020. Sentiment analysis of sinhala news comments using sentence-state lstm networks. pages 283–288.

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.

Lea Dieudonat, Kelvin Han, Phyllicia Leavitt, and Esteban Marquer. 2020. Exploring the combination of contextual word embeddings and knowledge graph embeddings. *ArXiv*, abs/2004.08371.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Philipp Dufter, Nora Kassner, and Hinrich Schütze. 2021. Static embeddings as efficient knowledge bases? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2353–2363, Online. Association for Computational Linguistics.

Carl Eckart and Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Luis Espinosa-Anke, Geraint Palmer, Padraig Corcoran, Maxim Filimonov, Irena Spasić, and Dawn Knight. 2021. English–welsh cross-lingual embeddings. *Applied Sciences*, 11(14):6541.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

Daniel C. Ferreira, André F. T. Martins, and Mariana S. C. Almeida. 2016. Jointly learning to embed and predict with multiple languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2019–2028, Berlin, Germany. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018a. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018b. Learning word vectors for 157 languages.

Prakhar Gupta and Martin Jaggi. 2021. Obtaining better static word embeddings using contextual embedding models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5241–5253, Online. Association for Computational Linguistics.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual LLMs to low-resource languages with knowledge graphs via adapters. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 63–74, Bangkok, Thailand. Association for Computational Linguistics.

Tim Isbister, Fredrik Carlsson, and Magnus Sahlgren. 2021. Should we stop training more monolingual models, and simply use machine translation instead? In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 385–390, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Dame Jovanoski, Veno Pachovski, and Preslav Nakov. 2015. Sentiment analysis in Twitter for Macedonian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 249–257, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Muhammad Yaseen Khan, Shah Muhammad Emaduddin, and Khurum Nazir Junejo. 2017. Harnessing english sentiment lexicons for polarity detection in urdu tweets: A baseline approach. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 242–249. IEEE.

Muhammad Yaseen Khan and Muhammad Suffian Nizami. 2020. Urdu sentiment corpus (v1.0): Linguistic exploration and visualization of labeled datasetfor urdu sentiment analysis. In *2020 IEEE 2nd International Conference On Information Science & Communication Technology (ICISCT)*. IEEE.

Elmurod Kuriyozov, Sanatbek Matlatipov, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2019. Construction and evaluation of sentiment datasets for low-resource languages: The case of uzbek. In *Human Language Technology. Challenges for Computer Science and Linguistics - 9th Language and Technology Conference, LTC 2019, Poznan, Poland, May 17-19, 2019, Revised Selected Papers*, volume 13212 of *Lecture Notes in Computer Science*, pages 232–243. Springer.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Siyu Li, Kui Zhao, Jin Yang, Xinyun Jiang, Zhengji Li, and Zicheng Ma. 2022. Senti-exlm: Uyghur enhanced sentiment analysis model based on xlm. *Electronics Letters*, 58(13):517–519.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

LocalDoc. 2024. Sentiment analysis dataset for Azerbaijani.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and R. Mamidi. 2022a. Multi-task text classification using graph convolutional networks for large-scale low resource language. *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022b. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–34.

Antonio Martínez-García, Toni Badia, and Jeremy Barnes. 2021. Evaluating morphological typology in zero-shot cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3136–3153, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *ArXiv*, abs/1309.4168.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013c. Exploiting similarities among languages for machine translation. *ArXiv*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013d. Distributed representations of words and phrases and their compositionality.

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

Shamsuddeen Muhammad, Idris Abdulmumin, Abinew Ayele, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard Opoku, and Stephen Arthur. 2023a. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023b. SemEval-2023 task 12: Sentiment analysis for African languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2319–2337, Toronto, Canada. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Samuel Pecar, Marian Simko, and Maria Bielikova. 2019. Improving sentiment classification in Slovak language. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 114–119, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3cube-mahasent-md: A multi-domain marathi sentiment analysis dataset and transformer models. In *Pacific Asia Conference on Language, Information and Computation*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Flora Sakketou and Nicholas Ampazis. 2020. A constrained optimization algorithm for learning glove embeddings with semantic lexicons. *Knowledge-Based Systems*, 195:105628.

Salim Sazzed. 2020. Cross-lingual sentiment classification in low-resource Bengali language. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 50–60, Online. Association for Computational Linguistics.

Oyesh Mann Singh, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi. 2020. Aspect based abusive sentiment detection in nepali social media texts. *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 301–308.

Charles Spearman. 1961. The proof and measurement of association between two things.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Robyn Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada. Association for Computational Linguistics.

Uga Sproģis and Matīss Rikters. 2020. What Can We Learn From Almost a Decade of Food Tweets. In *In Proceedings of the 9th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2020)*, Kaunas, Lithuania.

Nicolas Stefanovitch, Jakub Piskorski, and Sopho Kharazi. 2022. Resources and experiments on sentiment classification for Georgian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1613–1621, Marseille, France. European Language Resources Association.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Anca Tache, Gaman Mihaela, and Radu Tudor Ionescu. 2021. Clustering word embeddings with self-organizing maps. application on LaRoSeDa - a large Romanian sentiment data set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 949–956, Online. Association for Computational Linguistics.

Tarikwa Tesfa, Befikadu Belete, Samuel Abera, Sudhir Kumar Mohapatra, and Tapan Kumar Das. 2024. Aspect-based sentiment analysis on amharic text for evaluating ethio-telecom services. In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, pages 1–6.

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

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020b. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Ivan Vykopal, Simon Ostermann, and Marián Šimko. 2024. Soft language prompts for language transfer. In *2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics, Volume 1 (Long Papers)*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Fengqi Yan, Qiaoqing Fan, and Mingming Lu. 2018. Improving semantic similarity retrieval with word embeddings. *Concurrency and Computation: Practice and Experience*, 30(23):e4489.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Appendix

A Language Details

ISO code	Language	Size	Class	ConceptNet	ISO code	Language	Size	Class	ConceptNet
ss	Swati	86K	1	✘	**sc**	Sardinian	143K	1	✓
yo	Yoruba	1.1M	2	✓	**gn**	Guarani	1.5M	1	✓
qu	Quechua	1.5M	1	✓	**ns**	Northern Sotho	1.8M	1	✘
li	Limburgish	2.2M	1	✓	**ln**	Lingala	2.3M	1	✓
wo	Wolof	3.6M	2	✓	**zu**	Zulu	4.3M	2	✓
rm	Romansh	4.8M	1	✓	**ig**	Igbo	6.6M	1	✘
lg	Ganda	7.3M	1	✘	**as**	Assamese	7.6M	1	✘
tn	Tswana	8.0M	2	✘	**ht**	Haitian	9.1M	2	✓
om	Oromo	11M	1	✘	**su**	Sundanese	15M	1	✓
bs	Bosnian	18M	3	✘	**br**	Breton	21M	1	✓
gd	Scottish Gaelic	22M	1	✓	**xh**	Xhosa	25M	2	✓
mg	Malagasy	29M	1	✓	**jv**	Javanese	37M	1	✓
fy	Frisian	38M	0	✓	**sa**	Sanskrit	44M	2	✓
my	Burmese	46M	1	✓	**ug**	Uyghur	46M	1	✓
yi	Yiddish	51M	1	✓	**or**	Oriya	56M	1	✓
ha	Hausa	61M	2	✓	**la**	Lao	63M	2	✓
sd	Sindhi	67M	1	✓	**ta_rom**	Tamil Romanized	68M	3	✘
so	Somali	78M	1	✓	**te_rom**	Telugu Romanized	79M	1	✘
ku	Kurdish	90M	0	✓	**pu/pa**	Punjabi	90M	2	✓
ps	Pashto	107M	1	✓	**ga**	Irish	108M	2	✓
am	Amharic	133M	2	✓	**ur_rom**	Urdu Romanized	141M	3	✘
km	Khmer	153M	1	✓	**uz**	Uzbek	155M	3	✓
bn_rom	Bengali Romanized	164M	3	✘	**ky**	Kyrgyz	173M	3	✓
my_zaw	Burmese (Zawgyi)	178M	1	✘	**cy**	Welsh	179M	1	✓
gu	Gujarati	242M	1	✓	**eo**	Esperanto	250M	1	✓
af	Afrikaans	305M	3	✓	**sw**	Swahili	332M	2	✓
mr	Marathi	334M	2	✓	**kn**	Kannada	360M	1	✓
ne	Nepali	393M	1	✓	**mn**	Mongolian	397M	1	✓
si	Sinhala	452M	0	✓	**te**	Telugu	536M	1	✓
la	Latin	609M	3	✓	**be**	Belarussian	692M	3	✓
tl	Tagalog	701M	3	✘	**mk**	Macedonian	706M	1	✓
gl	Galician	708M	3	✓	**hy**	Armenian	776M	1	✓
is	Icelandic	779M	2	✓	**ml**	Malayalam	831M	1	✓
bn	Bengali	860M	3	✓	**ur**	Urdu	884M	3	✓
kk	Kazakh	889M	3	✓	**ka**	Georgian	1.1G	3	✓
az	Azerbaijani	1.3G	1	✓	**sq**	Albanian	1.3G	1	✓
ta	Tamil	1.3G	3	✓	**et**	Estonian	1.7G	3	✓
lv	Latvian	2.1G	3	✓	**ms**	Malay	2.1G	3	✓
sl	Slovenian	2.8G	3	✓	**lt**	Lithuanian	3.4G	3	✓
he	Hebrew	6.1G	3	✓	**sk**	Slovak	6.1G	3	✓
el	Greek	7.4G	3	✓	**th**	Thai	8.7G	3	✓
bg	Bulgarian	9.3G	3	✓	**da**	Danish	12G	3	✓
uk	Ukrainian	14G	3	✓	**ro**	Romanian	16G	3	✓
id	Indonesian	36G	3	✘					

Table 6: Details of the reproduced CC-100 corpus available on HuggingFace, including languages with their ISO codes, data set sizes, low-resource classifications, and language availability in the ConceptNet knowledge graph.

B SA Data Details

Language	ISO code	Source	#pos	#neg	#train	#val	#test
Sundanese	su	Winata et al., 2023	378	383	381	76	304
Amharic	am	Tesfa et al., 2024	487	526	709	152	152
Swahili	sw	Muhammad et al., 2023a; Muhammad et al., 2023b	908	319	738	185	304
Georgian	ka	Stefanovitch et al., 2022	765	765	1080	120	330
Nepali	ne	Singh et al., 2020	680	1019	1189	255	255
Uyghur	ug	Li et al., 2022	2450	353	1962	311	530
Latvian	lv	Sproģis and Rikters, 2020	1796	1380	2408	268	500
Slovak	sk	Pecar et al., 2019	4393	731	3560	522	1042
Sinhala	si	Demotte et al., 2020	2487	2516	3502	750	751
Slovenian	sl	Bučar et al., 2018	1665	3337	3501	750	751
Uzbek	uz	Kuriyozov et al., 2019	3042	1634	3273	701	702
Bulgarian	bg	Martínez-García et al., 2021	6652	1271	5412	838	1673
Yoruba	yo	Muhammad et al., 2023a; Muhammad et al., 2023b	6344	3296	5414	1327	2899
Urdu	ur	Maas et al., 2011; Khan et al., 2017; Khan and Nizami, 2020	5562	5417	7356	1812	1812
Macedonian	mk	Jovanoski et al., 2015	3041	5184	6557	729	939
Danish	da	Isbister et al., 2021	5000	5000	7000	1500	1500
Marathi	mr	Pingle et al., 2023	5000	5000	8000	1000	1000
Bengali	bn	Sazzed, 2020	8500	3307	8264	1771	1772
Hebrew	he	Amram et al., 2018	8497	3911	8932	993	2483
Romanian	ro	Tache et al., 2021	7500	7500	10800	1200	3000
Telugu	te	Marreddy et al., 2022b; Marreddy et al., 2022a	9488	6746	11386	1634	3214
Welsh	cy	Espinosa-Anke et al., 2021	12500	12500	17500	3750	3750
Azerbaijani	az	LocalDoc, 2024	14000	14000	19600	4200	4200

Table 7: Sentiment Analysis Data Details

C Common Vocabulary Counts

ISO code	GloVe and PPMI (Single)	GloVe and PPMI (All)
af	9,177	85,270
am	1,105	14,217
az	7,215	80,761
be	7,623	73,750
bn	3,962	38,221
bg	92,436	368,232
ku	3,762	32,499
cy	7,774	57,522
da	38,095	450,290
el	19,710	197,647
eo	59,476	161,634
et	14,815	163,666
gd	6,415	24,430
ga	13,871	65,169
gl	29,654	215,868
gu	3,198	24,575
ht	1,557	13,304
ha	671	33,824
he	16,032	153,731
hy	14,951	60,756
is	27,007	143,567
ja	2,607	41,471
kn	2,181	24,783
ka	17,869	96,066
kk	8,292	64,494
km	2,654	34,014
ky	2,234	29,915
lo	269,010	373,012
lt	12,485	200,404
lv	17,450	183,088
ml	4,092	38,864
mr	3,211	33,552
mk	21,692	93,121
my	3,189	24,319
ne	2,650	21,479
pa	2,282	16,068
ps	847	15,904
ro	25,704	366,809
sa	3,336	12,101
si	943	27,536
sk	14,694	268,576
sl	45,153	229,429
so	533	18,088
su	1,236	26,068
sw	6,425	59,906
ta	4,596	60,906
tl	12,563	42,653
ug	764	4,798
uk	16,397	327,563
ur	4,662	44,530
uz	3,229	37,704
xh	1,650	15,709
yi	5,177	18,572
ms	34,022	152,500
yo	558	5,254
qu	2,056	11,046
wo	999	18,509
th	45,975	238,502

Table 8: Common Vocabulary between GloVe and PPMI Embedding Spaces

D Vocabulary Coverage

ISO code	SA		SIB		XNLI		MultiSimLex	
	G (%)	F (%)	G (%)	F (%)	G (%)	F (%)	G (%)	F (%)
am	78.22	99.48	84.36	99.73	–	–	–	–
su	78.66	99.92	79.39	99.94	–	–	–	–
sw	88.24	100.00	91.32	99.98	83.68	99.98	73.94	100.00
si	89.18	99.99	91.63	99.97	–	–	–	–
ka	97.19	99.99	94.71	100.00	–	–	–	–
ne	77.91	99.82	84.93	99.99	–	–	–	–
ug	88.28	99.92	82.87	99.96	–	–	–	–
yo	22.37	99.18	46.50	99.73	–	–	–	–
ur	62.54	99.72	92.97	99.95	73.42	99.86	–	–
mk	82.90	99.92	95.84	99.99	–	–	–	–
mr	84.06	99.94	87.13	99.99	–	–	–	–
bn	66.55	99.75	89.58	100.00	–	–	–	–
te	85.66	99.99	–	–	–	–	–	–
uz	71.17	99.94	83.61	99.99	–	–	–	–
az	60.60	100.00	94.03	100.00	–	–	–	–
bg	84.18	99.91	98.16	100.00	96.47	99.98	–	–
sl	91.79	100.00	97.92	100.00	–	–	–	–
lv	87.04	99.41	97.43	99.97	–	–	–	–
sk	84.74	99.75	98.29	99.99	–	–	–	–
ro	90.16	99.94	98.71	100.00	–	–	–	–
he	89.72	99.74	97.57	100.00	–	–	91.79	100.00
cy	51.87	99.91	90.76	99.98	–	–	82.73	100.00
da	75.48	99.71	96.76	100.00	–	–	–	–
el	–	–	98.15	99.94	97.34	100.00	–	–
th	–	–	–	–	22.29	100.00	–	–
af	–	–	90.05	99.95	–	–	–	–
be	–	–	94.59	99.95	–	–	–	–
eo	–	–	93.83	100.00	–	–	–	–
et	–	–	94.50	100.00	–	–	94.70	99.99
gd	–	–	70.48	99.85	–	–	–	–
ga	–	–	89.97	99.93	–	–	–	–
gl	–	–	97.33	99.98	–	–	–	–
gu	–	–	87.11	99.97	–	–	–	–
ht	–	–	75.89	99.74	–	–	–	–
ha	–	–	87.20	–	–	–	–	–
hy	–	–	92.70	99.92	–	–	–	–
is	–	–	92.22	99.94	–	–	–	–
ja	–	–	82.97	99.98	–	–	–	–
kn	–	–	86.82	100.00	–	–	–	–
kk	–	–	93.11	100.00	–	–	–	–
km	–	–	24.06	99.92	–	–	–	–
ky	–	–	87.29	100.00	–	–	–	–
lo	–	–	19.17	–	–	–	–	–
lt	–	–	97.70	100.00	–	–	–	–
ml	–	–	85.13	100.00	–	–	–	–
my	–	–	31.18	99.96	–	–	–	–
pa	–	–	85.15	99.98	–	–	–	–
ps	–	–	78.72	99.91	–	–	–	–
sa	–	–	46.87	99.94	–	–	–	–
so	–	–	79.48	99.89	–	–	–	–
tl	–	–	89.07	100.00	–	–	–	–
uk	–	–	97.72	100.00	–	–	–	–
xh	–	–	62.39	–	–	–	–	–
yi	–	–	73.63	99.87	–	–	–	–
ms	–	–	95.71	100.00	–	–	–	–
qu	–	–	36.33	99.96	–	–	–	–
wo	–	–	54.05	–	–	–	–	–

Table 9: Vocabulary Coverage by GloVe FastText Embeddings for 4 Evaluation Tasks - Sentiment Analysis, Topic Classification, Natural Language Inference, and MultiSimLex

E Correlation Between Improvement Scores and Vocabulary Overlap

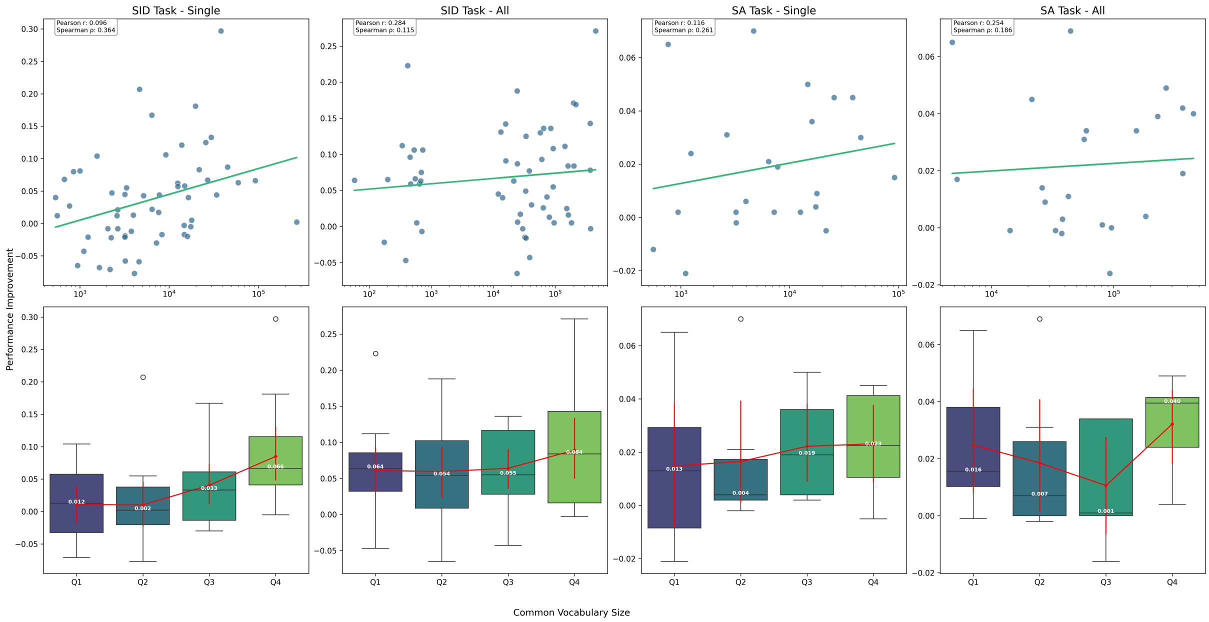

Figure 1: Scatter plots illustrating the relationship between vocabulary overlap and performance improvements across various language tasks using GloVe and graph-enhanced embeddings (G+P). Each plot shows the improvement in performance (G+P − GloVe) versus the common vocabulary size (log-scaled). Solid lines represent the best-fit log-linear trend.