Linear Cross-Lingual Mapping of Sentence Embeddings

Oleg Vasilyev, Fumika Isono, John Bohannon
Primer Technologies Inc.
San Francisco, California
oleg,fumika.isono,john@primer.ai

Abstract

Semantics of a sentence is defined with much less ambiguity than semantics of a single word, and we assume that it should be better preserved by translation to another language. If multilingual sentence embeddings intend to represent sentence semantics, then the similarity between embeddings of any two sentences must be invariant with respect to translation. Based on this suggestion, we consider a simple linear cross-lingual mapping as a possible improvement of the multilingual embeddings. We also consider deviation from orthogonality conditions as a measure of deficiency of the embeddings.

1 Introduction

The approximately linear mapping between cross-lingual word embeddings in different languages is based on assumption that the word semantic meaning is conserved in a translation (Mikolov et al., 2013). The linearity is only approximate because the corresponding words in different languages have different cultural background, different multiple meanings and different dependencies on context (Patra et al., 2019; Zhao and Gilman, 2020; Cao et al., 2020; Peng et al., 2020). There are multiple patterns of polysemy, and the corresponding counts of word senses are different in different languages (Srinivasan and Rabagliati, 2015; Casas et al., 2019).

We expect, however, that a sentence has a less ambiguous meaning than a word, simply because the sentence context reduces ambiguity of each of its words. Indeed, in (Kang et al., 2024) it is demonstrated that additional context helps to reduce disambiguation errors. The idea that a sentence semantics should be better conserved in a translation was used in (Reimers and Gurevych, 2020).

In Appendix A we provide simple examples illustrating the loss of word ambiguity in a sentence, and suggest that a good translation can preserve the residual ambiguity, if any. The examples show that if semantics of a sentence is somewhat changed in translation, then a better translation is possible. Unlike a lone word, which often has different sets of meaning in different languages, a sentence is not only less ambiguous but also allow differently phrased translations, among which there is usually at least one that fully preserves the semantics.

In order to explore the preservation of sentence semantics in translation, we consider here a linear mapping between multilingual embeddings in two languages. Unlike the removal of a language-specific bias in each language separately (Yang et al., 2021; Xie et al., 2022), this mapping depends on both languages of interest and, while computationally cheap, may provide a better correspondence between the embeddings. Our contribution:

1. We suggest simple and computationally light improvement of the correspondence of sentence embeddings between two languages. The 'sentence' can be one or several contiguous sentences.
2. For our evaluation we introduce a dataset based on wikipedia news.
3. We demonstrate a non-orthogonality of the linear mapping between multilingual embeddings as an example and a measure of deficiency of a multilingual embedding model.

2 Cross-Lingual Linear Mapping

Translation of a word can lose or add some of its meanings. But meaning of a sentence or of several contiguous sentences is better defined, and a good translation in most cases (except special idiomatic cases) should preserve the semantics (Appendix A). Embeddings of the translated sentences should be rigidly related to embeddings of the original sentences: the semantic similarities (or distances) between different embeddings should be preserved.

Findings of the Association for Computational Linguistics ACL 2024, pages 8163–8171
August 11-16, 2024 ©2024 Association for Computational Linguistics

In this section we assume that the 'sentence' is either a (not too short) sentence, or a larger segment of a text.

Suppose we have N sentences, translated from language L to language L', and then embedded into a space of the same dimension M in each of these languages: the embeddings $e_1, ...e_N$ in L and the embeddings $e'_1, ...e'_N$ in L'. If the measure of semantic similarity in both spaces is cosine, then we should expect that the normalized embeddings e_i and e'_i are related by rotation (orthonormal transform T):

$$e' = Te \tag{1}$$

with the orthogonality condition

$$\sum_i T_{ij} T_{ik} = \delta_{jk} \tag{2}$$

where $i, j, k = 0, 1, ..., M-1$.

If semantic similarity is measured by euclidean distance, and the embeddings are not normalized, then we should allow the orthogonal transform to be accompanied by dilation and shift:

$$e' = \alpha Te + b \tag{3}$$

The above transformations should be observed if the translations preserved the semantics of the sentences, and if the embeddings represent the semantics correctly.

In the following section we will allow any linear transformation (A, b) between the embeddings in L and L':

$$\tilde{e} = Ae + b \tag{4}$$

For our illustration here we created embeddings by one of SOTA aligned multilingual sentence-embedding model, on a set of translated sentences (Section 3.2). We optimize the linear transformation on a set of embeddings, so that the mean squared distance between $\tilde{e}$ and e' is minimal.

In the next section we consider the obtained linear transformation (A, b) from two points of view:

1. Replacement of the original embeddings e by the transformed embeddings $\tilde{e}$ can serve as a fast and computationally cheap way to improve cross-lingual matching or clustering of a mix of texts of both languages.
2. We can observe how close is the optimized transformation (A, b) to the 'ideal' relation eq.3, and thus judge how good the embeddings are.

3 Observations

3.1 Data

For obtaining the linear transformation eq.4 between embeddings, in Section 3.2 we use dataset Tatoeba[1]. Tatoeba has 13 languages with at least $100K$ sentences translated from English to the language. We consider performance of the obtained transformations on sentences and text segments of different style from multilingual WikiNews dataset[2] which we created from real news (Appendix B). The samples have WikiNews articles in English as well as at least one other language, among 34 languages.

We will limit ourselves to six languages L' that have a reasonable amount of data: at least $100K$ samples (of translations from L to English) in Tatoeba, and at least 400 samples in Wikinews (Appendix B): German (de), Spanish (es), French (fr), Italian (it), Portuguese (pt) and Russian (ru). Wikinews is used here for evaluation, in Section 3.2, in two variations:

1. *WN*: Title of news article in English is paired with the same title in language L'.
2. *WN-text*: Title of news article in English is paired with the lower half of the text of the article in language L'. We selected the lower part in order to avoid easy lexical intersections of first phrases of the text with the title. (The article is split by whichever end of sentence is closer to the middle.)

The evaluation on title-title pairs gives us a strong out-of-domain experience, and evaluation on title-text pairs provides a (more difficult) flavor of asymmetry in a multilingual search. We also evaluate the obtained transformations on Flores dataset (Guzmán et al., 2019; Goyal et al., 2022; Team et al., 2022)[3], and on a Tatoeba subset left aside from training.

3.2 Evaluation

We obtained the transformation (A, b) (eq.4) for each language $L' = de, es, fr, it, pt, ru$ by (1) obtaining embeddings e for English sentences and embeddings e' for the sentence translations to language L', and (2) training a simple linear layer with bias, using embeddings e_i as the inputs, and embeddings e'_i as the labels, with the distance $|\tilde{e}_i - e'_i|$

[1] https://huggingface.co/datasets/tatoeba

[2] https://huggingface.co//datasets//Fumika//Wikinews-multilingual

[3] https://huggingface.co/datasets/facebook/flores

serving as loss function. For each language, $10K$ embedding pairs were set aside for the testing, and $10K$ embedding pairs were set aside and used for validation during the training. We used state of the art embeddings paraphrase-multilingual-mpnet-base-v2 (Reimers and Gurevych, 2019)[4] for obtaining the embeddings e and e'.

We can evaluate the benefit of replacing the original embeddings e by the transformed embeddings $\tilde{e}$ in different ways. In Table 1 we consider several examples: dD, dC, fD, fC - defined below.

data	lang	dD	dC	fD	fC
Tatoeba	de	0.152	0.009	0.709	0.670
	es	0.081	0.003	0.686	0.566
	fr	0.124	0.007	0.688	0.650
	it	0.084	0.005	0.604	0.562
	pt	0.078	0.004	0.637	0.605
	ru	0.129	0.006	0.712	0.678
WN	de	0.133	0.016	0.954	0.827
	es	0.063	0.002	0.936	0.618
	fr	0.114	0.012	0.938	0.799
	it	0.085	0.009	0.961	0.785
	pt	0.075	0.005	0.848	0.630
	ru	0.167	0.024	0.982	0.890
WN-text	de	0.182	0.039	1.000	0.988
	es	0.082	0.008	1.000	0.912
	fr	0.144	0.030	1.000	0.981
	it	0.111	0.021	1.000	0.963
	pt	0.118	0.023	1.000	0.963
	ru	0.192	0.037	1.000	0.991
Flores	de	0.084	0.002	0.709	0.502
	es	0.066	0.000	0.914	0.502
	fr	0.084	0.002	0.746	0.526
	it	0.069	0.001	0.820	0.494
	pt	0.053	0.001	0.713	0.502
	ru	0.186	0.006	0.926	0.696

Table 1: Performance of the linear transform $e \to \tilde{e}$ (eq.4), trained on Tatoeba dataset, and evaluated on (set aside) Tatoeba, WN (Wiki-news title-to-title), WN-text (Wiki-news title-to-halftext), and Flores. Performance is estimated as improvement in average distance dD (eq.5) and in average cosine dC (eq.8), fraction of samples with improved distance fD (eq.9), and fraction of samples with improved cosine fC (eq.10).

The measure

$$dD = \frac{d - \tilde{d}}{\min{(d, \tilde{d})}} \tag{5}$$

compares the achieved average distance

$$\tilde{d} = \frac{1}{N} \sum_{i}^{N} |\tilde{e}_i - e'_i| \tag{6}$$

and the original distance

$$d = \frac{1}{N} \sum_{i}^{N} |e_i - e'_i| \tag{7}$$

where the embeddings e are taken for a test dataset of size N. The measure

$$dC = \frac{1}{N} \sum_{i}^{N} \left(\cos(\tilde{e}_i, e'_i) - \cos(e_i, e'_i) \right) \tag{8}$$

compares the cosines. It is similar to comparing distances in eq. 5; there is no need here for normalization, and the improvement is measured by the increase of cosine (whereas in eq. 5 it was the decrease of distance).

While on average the alignment of the embeddings may improve (as indeed is the case in our evaluations, showing dD and dC being positive in Table 1), the improvement is not evenly distributed between the samples. We would like to assess how many samples benefit from the transformation. The measure

$$fD = \frac{1}{N} \sum_{i}^{N} \left(H(|e_i - e'_i| - |\tilde{e}_i - e'_i|) \right) \tag{9}$$

where H is the Heaviside step function, represents the fraction of the samples for which the distance has decreased.

Similarly, the measure

$$fC = \frac{1}{N} \sum_{i}^{N} \left(H(\cos(\tilde{e}_i, e'_i) - \cos(e_i, e'_i)) \right) \tag{10}$$

represents the fraction of the samples for which the cosine increased.

The transformation $e \to \tilde{e}$ helps if dD and dC are positive (the higher the better), and if the fractions fD and fC are higher than 0.5 (the higher the better, for having an improvement in the majority of samples). The measures dD and fD should be of interest when matching of embeddings (e.g. search) is to be done by distance; the measures dC and fC are of interest for matching by cosine. Table 1 shows that these conditions are satisfied for almost all cases. The only exception is the value of fC for Italian (it) language in Flores dataset: here the cosine got improved for slightly less that half (49.4%) of the samples.

[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

3.3 Orthogonality

If a good translation indeed fully preserves the semantics of a sentence, and if the embedding model would produce ideal alignment, then the sentence embeddings in different languages would be close to identical: $e' = e$. The transform T (eq.1) would then become an identity. If the embedding model does not perfectly align the embeddings e and e' (or does not align them at all), but still correctly embed their semantics in each of the languages L and L', then the optimized linear transformation (A, b) (eq.4) must be orthogonal as in eq.3.

In order to evaluate how close our linear transformation A (trained on Tatoeba) to being orthogonal (Eq.2), we consider the values

$$p_{jk} = \frac{\sum_i A_{ij} A_{ik}}{|A_j| \cdot |A_k|} \ , \qquad j \neq k \tag{11}$$

where

$$|A_j| = \sqrt{\sum_i A_{ij}^2} \tag{12}$$

The closer these values p_{jk} to zero, the closer A to being orthogonal. In Table 2 we show simple aggregates of p_{ij} over all $i \neq j$. The measure $\langle|p|\rangle$ is an average of absolute values of non-diagonal elements:

$$\langle|p|\rangle = \frac{1}{M(M-1)} \sum_{j \neq k} |p_{jk}| \tag{13}$$

where M is the dimensionality of the embeddings $(j, k = 0, 1, ..., M-1)$.

The orthogonality may be compromised for some embeddings more than for others. To characterise this, we show in Table 2 the standard deviation

$$\sigma(p) = \sqrt{\frac{1}{M(M-1)} \sum_{j \neq k} (p_{jk} - \langle p \rangle)^2} \tag{14}$$

where the average $\langle p \rangle$ is

$$\langle p \rangle = \frac{1}{M(M-1)} \sum_{j \neq k} p_{jk} \tag{15}$$

We show also $\min(p)$ and $\max(p)$:

$$\min(p) = \min_{j \neq k} p_{jk} \qquad \max(p) = \max_{j \neq k} p_{jk} \tag{16}$$

Table 2 lists more languages than Table 1 because there is no need here to apply A to other datasets: we are simply considering the orthogonality of A.

The highest by far deviation from orthogonality in Table 2 is for Berber (ber) language, followed by Esperanto (eo). The minimal and maximal values are colored yellow when they exceed 0.383, meaning that for at least one pair i, j the angle is less than 75% of orthogonal $(\cos(\pi/2 * 0.75) \approx 0.383$.

lang	$\langle\lvert p\rvert\rangle$	$\sigma(p)$	$\min(p)$	$\max(p)$
ber	0.204	0.254	-0.861	0.845
de	0.019	0.025	-0.154	0.337
eo	0.059	0.074	-0.362	0.345
es	0.004	0.005	-0.035	0.038
fr	0.019	0.024	-0.194	0.397
he	0.027	0.034	-0.353	0.516
it	0.011	0.014	-0.071	0.071
ja	0.032	0.042	-0.360	0.623
pt	0.013	0.017	-0.100	0.135
ru	0.018	0.023	-0.150	0.219
tr	0.027	0.035	-0.322	0.498
uk	0.020	0.026	-0.191	0.281

Table 2: Aggregates over orthogonality conditions Eq.11 for A trained on Tatoeba dataset, for languages containing at least $100K$ samples. Min and max beyond 25% deviation from orthogonality $(\cos(0.75\pi/2) \approx 0.383)$ are colored yellow.

For comparison, in Table 3 we show similar data for A trained on United Nations Parallel Corpus UNPC (Ziemski et al., 2016)[5] (with 500K samples used for training and 10K for validation). The UN texts have a specific formal style and meant to be precise in dealing with loaded topics. The translations are also intended to be precise, conserving semantics. But these documents' cumbersome formal style and some very long sentences may be more difficult than the common texts for an embedding model. Indeed, for each of the three languages common for Tatoeba Table 2 and UNPC Table 3 (Spanish es, French fr and Russian ru) all the aggregate indicators $\langle|p|\rangle$, $\sigma(p)$, $\min(p)$ and $\max(p)$ are several times larger for UNPC-trained matrix A (Table 3).

The orthogonal transformation can be accompanied by dilation (coefficient α in Eq.3), which means that the values $\alpha_i = |A_i|$ (eq.12) should not depend on i. In order to assess deviations from this condition, we consider normalized standard deviation

$$\frac{\sigma(\alpha)}{\bar{\alpha}} = \frac{1}{\bar{\alpha}} \sqrt{\frac{1}{M} \sum_i (\alpha_i - \bar{\alpha})^2} \tag{17}$$

[5] https://conferences.unite.un.org/uncorpus

lang	$\langle \lvert p \rvert \rangle$	$\sigma(p)$	$\min(p)$	$\max(p)$
ar	0.026	0.033	-0.147	0.157
es	0.014	0.018	-0.130	0.107
fr	0.144	0.195	-0.769	0.795
ru	0.404	0.476	-0.958	0.950
zh	0.039	0.050	-0.254	0.495

Table 3: Aggregates over orthogonality conditions Eq.11 for A trained on UNPC.

and normalized range

$$r(\alpha) = \frac{\max \alpha - \min \alpha}{\bar{\alpha}} \tag{18}$$

where

$$\bar{\alpha} = \frac{1}{M} \sum_i \alpha_i \tag{19}$$

$$\min(\alpha) = \min_i \alpha_i \qquad \max(\alpha) = \max_i \alpha_i \tag{20}$$

The dilation quality measures $\frac{\sigma(\alpha)}{\bar{\alpha}}$ and $r(\alpha)$ are shown in Tables 4 and 5, for the transformations obtained on Tatoeba and on UNPC datasets correspondingly. The tables contain also the values of $\bar{\alpha}$ - the averaged α, and of the minimal and maximal values of α.

lang	$\bar{\alpha}$	$\frac{\sigma(\alpha)}{\bar{\alpha}}$	$r(\alpha)$	$\min(\alpha)$	$\max(\alpha)$
ber	0.637	0.336	1.856	0.258	1.440
de	0.814	0.039	0.275	0.753	0.977
eo	0.640	0.192	1.056	0.377	1.053
es	0.964	0.005	0.050	0.951	1.000
fr	0.845	0.034	0.230	0.791	0.986
he	0.814	0.060	0.333	0.727	0.998
it	0.889	0.021	0.170	0.841	0.992
ja	0.809	0.073	0.419	0.705	1.044
pt	0.877	0.022	0.174	0.838	0.990
ru	0.836	0.031	0.224	0.789	0.976
tr	0.835	0.054	0.307	0.751	1.007
uk	0.860	0.037	0.239	0.797	1.002

Table 4: Nonuniformity of dilation of embeddings transformation (Eqs.17, 18). For the transformation trained on Tatoeba dataset.

Similarly to the orthogonality conditions, the dilation quality measures $\frac{\sigma(\alpha)}{\bar{\alpha}}$ and $r(\alpha)$ are better (lower) for the transformation trained on Tatoeba (Table 4) than on UNPC (Table 5), for all three languages they have in common: Spanish (es), French (fr) and Russian (ru). Both measures generally follow similar trends across the languages.

lang	$\bar{\alpha}$	$\frac{\sigma(\alpha)}{\bar{\alpha}}$	$r(\alpha)$	$\min(\alpha)$	$\max(\alpha)$
ar	0.761	0.088	0.470	0.630	0.988
es	0.840	0.043	0.253	0.767	0.980
fr	0.938	0.190	1.126	0.700	1.756
ru	1.338	0.444	2.908	0.661	4.551
zh	0.865	0.114	0.559	0.696	1.180

Table 5: Nonuniformity of dilation of embeddings transformation (Eqs.17, 18). For the transformation trained on UNPC.

As we could already expect from observations in Table 2, the measures $\frac{\sigma(\alpha)}{\bar{\alpha}}$ and $r(\alpha)$ in Table 4 are the worst for Berber (ber) and Esperanto (eo) languages. A distant third (also as in Table 4) is Japanese language (ja).

For most languages the ratio $\frac{\sigma(\alpha)}{\bar{\alpha}}$ may look comfortably small, but the normalized range $r(\alpha)$ is high for some languages in both tables 4 and 5. Altogether, we have to conclude that orthogonality is only approximately satisfied by the linear transform (A, b).

4 Conclusion

We considered a simple and inexpensive method of improving the alignment between sentence embeddings in two languages: a linear transformation, tuned on embeddings of the paired sentences. In the examples we analyzed, a training on sentences also improves an alignment between titles and texts (lower-half texts) of the articles - the articles from our WikiNews dataset.

If embeddings were capable of perfectly encoding semantics even when not perfectly aligned, then the linear transformation would be an orthogonal transformation, accompanied by dilation and shift. Measuring deviation from this condition allows us to judge the quality of the embeddings. For example, we observed lower quality for embeddings of Berber and Esperanto languages compared to other languages considered here, and also a lower quality of UNPC-trained transformations compared to Tatoeba-trained transformations.

It would be interesting to consider deviation from orthogonality for individual samples, as the strong deviations could point either to bad translations or to the samples difficult to embed by the model.

Limitations

Our consideration involved a limited set of languages. This limitation allowed us to evaluate Tatoeba-trained transformations on very different styles of matching sentences, but the research can be extended to many more languages.

We suggested simple measures of quality of multilingual embeddings based on the orthogonality requirement (Section 3.3). While our observations confirm that these measures are reasonable, we do not claim that these are the best possible measures.

We have not considered a possibility of measuring the deviations from orthogonality by individual samples. If such samples are particularly imperfect translation (see Appendix A) then removing such samples from the dataset used for tuning would improve orthogonality of the transformation, and hence would make better the introduced here measures of the quality of embeddings.

A complementing possibility is that a very good embedding model could help to identify imperfect translations; this may be unlikely because the embeddings are very approximate in encoding the semantics, but we do not provide definitive observations.

The role of polysemy and its variation between languages is not investigated here beyond the intuitive arguments and examples of Appendix A, in which we suggest that in most cases the context removes or strongly reduces ambiguity, and a good translation keeps the residual ambiguity, if any, unchanged.

Acknowledgments

We thank Randy Sawaya for many discussions and review of the paper. We also thank an anonymous reviewer for concern about the polysemy problem.

References

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations.

Bernardino Casas, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer i Cancho, and Jaume Baixeries. 2019. Polysemy and brevity versus frequency in language. *Computer Speech Language*, 58:19–50.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Translate to disambiguate: Zero-shot multilingual word sense disambiguation with pretrained language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1562–1575, St. Julian's, Malta. Association for Computational Linguistics.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Xutan Peng, Mark Stevenson, Chenghua Lin, and Chen Li. 2020. Understanding linearity of cross-lingual word embedding mappings. *arXiv*.

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*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Mahesh Srinivasan and Hugh Rabagliati. 2015. How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua*, 157:124–152. Polysemy: Current Perspectives and Approaches.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. 2022. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5617–5633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve. 2021. A simple and effective method to eliminate the self language bias in multilingual representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5825–5832, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiawei Zhao and Andrew Gilman. 2020. Non-linearity in mapping based cross-lingual word embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3583–3589, Marseille, France. European Language Resources Association.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Loss of Ambiguity

A.1 Polysemy problem

As we discussed in Introduction (Section 1), we assume that ambiguity of words is mostly lost with context. On one hand, it is intuitively understandable, and the role of context was demonstrated in (Kang et al., 2024). On the other hand, polysemy is occasionally possible even with a context. A word ambiguity can be intentionally and skillfully kept through many sentences or a long dialog for sake of misinterpretation comedy. Also, the available word polysemy greatly varies across languages and patterns (see for example Table 5 in (Srinivasan and Rabagliati, 2015), or Table 3 in (Casas et al., 2019)).

Yet, the common sentences are remarkably unambiguous. For example, the word 'board' looses all or most of its multiple definitions in arbitrary sentences of lengths 3 to 7 words, generated by a GPT3.5 (our examples are in Table 6).

The board cracked.
The board is white.
The board is now full.
Circuit board malfunctioned, causing system failure.
The cork board holds important reminders daily.

Table 6: Examples of sentences with the word 'board'.

Even the residual ambiguity may be kept unchanged by a good translation. As a simple illustration that a typical sentence is loosing ambiguity of its words, and that the residual ambiguity is usually kept intact in a good translation, we examined (in the following subsections) the first 10 sentences from Tatoeba and from Flores datasets, reviewing the sentences in English, French, Japanese, Russian, Spanish and Ukrainian. (We also reviewed top 10 sentences from UNPS in English, French, Russian and Spanish, and could not find any change in semantics in those meticulous formal style translations.)

Despite many multi-sense words in all the sentences considered below, there were few examples where the semantics of English sentence would not exactly correspond to semantics of the translated sentence. The examples show that if semantics of a sentence is somewhat changed in translation, it is mostly due to a deficiency of translation rather than some impenetrable polysemy barrier between the languages.

A.2 Examples from Tatoeba

We have not found any shift of semantics in the first 10 samples from English-French part of Tatoeba. For example, the very first sample uses a few words that could be used in different senses, but the semantics of English "When he asked who had broken the window, all the boys put on an air of innocence." is well matched by French "Lorsqu'il a demandé qui avait cassé la fenêtre, tous les garçons ont pris un air innocent.".

In the case of the first 10 samples of English-Japanese pairs from Tatoeba, one example has a shift in semantics. The Japanese translation "ムーリエルは２０歳になりました。" means Muiriel has just turned 20 years old, whereas the English translation "Muiriel is 20 now." does not indicate whether Muiriel has just turned 20 or has been 20 for a while. Additionally, the over-restriction of the meaning of "20" to age, as seen on English-Ukrainian pairs, is also observed in the translation to Japanese.

Of the first 10 samples from English-Spanish part of Tatoeba, we found 2 samples where the semantics is shifted: The English sentence "Let's try something." is translated (in two out of four versions) using the word "permiteme", which narrows down the meaning by suggesting that it is the speaker that would "try something".

Of the first 10 samples from English-Ukrainian part of Tatoeba, we found 2 samples where semantics is somewhat shifted. One of three translations of the English sentence "I have to go to sleep." is over-specific “Мені час йти спати.”, narrowing the reason (time). Also, one of three translations of "Muiriel is 20 now." is "Мюріел зараз двадцять років.". Strictly speaking, the English sentence could also be used in a game or sport to inform about some score, while this particular translation to Ukrainian narrows down "20" as age.

Of the first 10 samples from English-Russian part of Tatoeba, there is one sample with shifted semantics: Similar to Ukrainian samples, one of three translations of "I have to go to sleep." is over-specific "Мне пора идти спать." (meaning "It is time for me to go to sleep.").

A.3 Examples from Flores

Sentences in Flores, unlike in Tatoeba, are long sentences like one would encounter in informative news. Each sample consists of an English sentence is translated to many languages. Of the first 10 samples examined for translation to French, Japanese, Russian, Spanish and Ukrainian languages, we could find only three examples of the translation changing semantics.

There are two examples in English-Japanese pairs where the Japanese translations differ semantically from the English ones. In sample #1, the English phrase "about one U.S. cent each" is translated to "1円ほどす。". First, there is a typographical error where "ほどす。" should be typed "ほどです。". This could result in misalignment of the semantics of the sentence pair. Secondly, there is a semantic shift in translation. Its literal translation is "about 1 yen", which uses Japanese currency. Although a cent and a yen are of similar value (currently, 1 cent is about 1.6 yen), changing the currency unit in translation can significantly alter the sentence's meaning. Another example is the English phrase "closing the airport to commercial flights" in sample #3. Its Japanese translation is "空港の商業便が閉鎖されました。", which literally means "Commercial flights in the airport were closed," where the object of the verb "close" is "flights," not the airport. Confusing the subject and object can change the semantic meaning of the sentences.

There is one example (sample #8) of semantics shift in English-Ukrainian pairs: In the translation of English sentence "The protest started around 11:00 local time (UTC+1) on Whitehall opposite the police-guarded entrance to Downing Street, the Prime Minister's official residence." to Ukrainian the word "біля" (meaning "near") was used for "opposite", thus adding a bit of ambiguity.

B WikiNews

The WikiNews dataset[6][7] comprises 15,200 news articles from the multilingual WikiNews website[8], including 9,960 non-English articles written in 33 different languages. These articles are linked to one of 5,240 sets of English news articles as WikiNews pages in other languages. Therefore, these WikiPages in different languages can be assumed to be describing the same news event, thus we can assume that the news titles and contents are of the linked NewsPages are semantically alligned. Here the non-English articles are written in a variety of languages including Spanish, French, German, Portuguese, Polish, Italian, Chinese, Russian, Japanese, Dutch, Swedish, Tamil, Serbian, Czech, Catalan, Hebrew, Turkish, Finnish, Esperanto, Greek, Hungarian, Ukrainian, Norwegian, Arabic, Persian, Korean, Romanian, Bulgarian, Bosnian, Limburgish, Albanian, and Thai.

Each sample in the multilingual WikiNews dataset includes several variables, such as pageid, title, categories, language, URL, article content, and the publish date. In some cases, foreign WikiNews sites may have news titles but no content, in which case the text variable is left empty. Samples with the same pageid in the dataset correspond to the same news event, which are linked together as the same WikiNews pages with other languages. The published date of an English sample is scraped and converted to DateTime format, but dates in foreign samples are left as is. Table 7 shows the example samples of the dataset.

The number of samples for the languages used in Table 1: de: 1053; es: 1439; fr: 1311; it: 618;

[6] https://huggingface.co//datasets//Fumika//Wikinews-multilingual

[7] https://github.com//PrimerAI//primer-research

[8] https://www.wikinews.org/

Table 7: Example of samples from the multilingual WikiNews dataset

index	**pageid**	**lang**	**title**	**content**
0	232226	en	“Very serious”: Chinese government releases corruption report	A report by the Chinese government states corruption is "very serious". ...
1	232226	cs	Čína připustila, že tamní korupce je vážný problém	Zpráva čínské vlády připouští, že korupce v zemi je stále „velmi vážná“, jelikož úřady ...
2	232226	es	China admite que la corrupción en el país es “muy seria”s	29 de diciembre de 2010Beijing, China — Un reporte del gobierno de la República Popular China ...

pt: 1023; ru: 436.