Appendix

Anonymous Author(s)
Affiliation
Address
email

A Training Corpora and Hyper-parameters

A.1 Training Corpora

As for monolingual data, we follow Conneau et al. [2020] to build a Common-Crawl Corpus using the CCNet [Wenzek et al., 2020] tool[1], which is widely used in the literature Huang et al. [2019], Luo et al. [2021], Chi et al. [2021], Wei et al. [2021]. Further, we collect parallel corpora from CCAligned El-Kishky et al. [2020], CCMatrix Schwenk et al. [2021], WMT Akhbardeh et al. [2021], and MultiUN Ziemski et al. [2016], involving 94 languages with more than 4.8 billion sentence pairs. We use the OpusFilter[2] tool to remove noisy bitexts, which results in 3.2 billion sentence pairs. Table 1 shows the statistics for both monolingual and parallel data. We apply subword tokenization directly on raw text data using Sentence Piece Model Kudo and Richardson [2018] without any additional preprocessing. To better support our motivation that EMMA-X can cover more languages than previous cross-lingual sentence representations, we divide Tatoeba Artetxe and Schwenk [2019] into two subsets: "Head", containing languages usually covered in previous methods, and "Long-tail", with other languages. We treat the 36 languages containing in XTREME Ruder et al. [2021] as head languages, which are: **"ar, he, vi, id, jv, tl, eu, ml, ta, te, af, nl, en, de, el, bn, hi, mr, ur, fa, fr, it, pt, es, bg, ru, ja, ka, ko, th, sw, zh, kk, tr, et, fi, hu, az, lt, pl, uk, ro"**. The remaining 76 languages in Tatoeba are treated as long-tail ones.

A.2 Hyper-parameters

The parameters of EMMA-X are first initialized with XLM-R, with 24 layers of Transformer [Vaswani et al., 2017] encoder, 1024 hidden states, and 16 attention heads. We set the total semantic ranks as 4. The GMM classifier is implemented as a mixture of Gaussian forms, each of which consists of a prior $\pi \in \mathbb{R}^1$, a mean $\mu \in \mathbb{R}^{1024}$, and a variance $\sigma \in \mathbb{R}^{1024}$, all are trainable variables. We optimize the GMM classifier with Adam (β_1=0.9, β_2=0.999) Kingma and Ba [2015] using a batch size of 1024 and a learning rate of 3e-5. For cross-lingual encoder, we apply the same training setting as MoCo He et al. [2020], with the momentum queue K to be 256 and temperature as 0.04. We set the momentum coefficient to 0.999 and use the Adam optimizer with a cosine decay learning rate whose peak is 5e-4.

B FLORES-200 Dataset and Geometric Analysis

B.1 FLORES-200 dataset

FLORES-200 Goyal et al. [2022], Costa-jussà et al. [2022] is a many-to-many multilingual benchmark, which consists of 3001 sentences in 204 total languages. FLORES-200 sourced all sentences from English WikiMedia and translated these English sentences to 204 languages by human translators. In particular, sentences in FLORES-200 have a much larger breadth of topics, for they are

[1] https://github.com/facebookresearch/cc_net
[2] https://github.com/Helsinki-NLP/OpusFilter

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Code	Size (GB)	Sent. (M)	Code	Size (GB)	Sent. (M)	Code	Size (GB)	Sent. (M)	Code	Size (GB)	Sent. (M)	Code	Size (GB)	Sent. (M)
af	1.3	-	et	6.1	22.3	ja	24.2	89.2	mt	0.2	-	sq	3.0	-
am	0.7	-	eu	2.0	0.81	jv	0.2	-	my	0.9	-	sr	5.1	-
ar	20.4	72.3	fa	21.6	7.5	ka	3.4	2.0	ne	2.6	-	su	0.1	-
as	0.1	-	fi	19.2	92.8	kk	2.6	2.8	nl	15.8	66.0	sv	10.8	74.2
az	3.6	0.82	fr	46.5	331.5	km	1.0	0.84	no	3.7	-	sw	1.6	1.7
be	3.5	0.51	fy	0.2	0.13	kn	1.2	-	om	0.1	-	ta	8.2	2.79
bg	22.6	47.2	ga	0.5	-	ko	17.2	79.3	or	0.6	-	te	2.6	-
bn	7.9	7.52	gd	0.1	0.05	ku	0.4	-	pa	0.8	-	th	14.7	13.1
br	0.1	-	gl	2.9	0.77	ky	1.2	-	pl	16.8	79.7	tl	0.8	-
bs	0.1	-	gu	0.3	-	la	2.5	-	ps	0.7	-	tr	17.3	93.8
ca	10.1	14.9	ha	0.3	-	lo	0.6	-	pt	15.9	247.6	ug	0.4	-
cs	16.3	108.4	he	6.7	47.1	lt	7.2	11.0	ro	8.6	60.4	uk	9.1	0.78
cy	0.8	-	hi	20.2	3.2	lv	6.4	0.37	ru	48.1	134.9	ur	5.0	1.15
da	15.2	8.0	hr	5.4	-	mg	0.2	-	sa	0.3	-	uz	0.7	-
de	46.3	283.4	hu	9.5	55.2	mk	1.9	-	sd	0.4	-	vi	44.6	15.3
el	29.3	95.1	hy	5.5	1.7	ml	4.3	1.07	si	2.1	0.60	xh	0.1	-
en	49.7	-	id	10.6	184.6	mn	1.7	0.19	sk	4.9	-	yi	0.3	-
eo	0.9	0.18	is	1.3	-	mr	1.3	-	sl	2.8	9.8	zh	36.8	379.4
es	44.6	279.6	it	19.8	179.3	ms	3.2	2.1	so	0.4	-	-	-	

Table 1: The statistics of CC-100 and the collected parallel corpora used for training. We report the list of 94 languages and include the size of the monolingual data (in GiB) and the number of sentence pairs (in Millions, which denotes the number of sentence pairs between the specific language and English) in parallel corpora for each language. "-" means the number of sentence pairs is less than 0.1 million.

Number of Sentences		3001
Average Words per Sentence		21
Number of Articles		842
Average Number of Sentences per Article		3.5
Domain	**Articles**	**Sentences**
WikiNews	309	993
WikiJunior	284	1006
WikiVoyage	249	1002
Sub-Topic	**Articles**	**Sentences**
Crime	155	313
Disasters	27	65
Entertainment	28	68
Geography	36	86
Health	27	67
Nature	17	45
Politics	171	341
Science	154	325
Sports	154	162
Travel	505	1529

Table 2: Basic Statistics of FLORES-200.

collected from three different sources: WikiNews[3], WikiJunior[4] and WikiVoyage[5]. We summarize the basic statistics of all languages in FLORES-200 in Table 2. Similar to Tatoeba [Artetxe and Schwenk, 2019], we treat English data "eng_Latn" as retrieval labels and report the retrieval accuracy using the same scripts as Tatoeba in XTREME [Ruder et al., 2021]. We set the 68 languages: **"bel_Cyrl, bos_Latn, hun_Latn, epo_Latn, khm_Khmr, urd_Arab, srp_Cyrl, jav_Latn, hye_Armn, gla_Latn, por_Latn, lit_Latn, bul_Cyrl, slk_Latn, mal_Mlym, ita_Latn, nno_Latn, mar_Deva, hrv_Latn, hin_Deva, kat_Geor, ben_Beng, fin_Latn, cym_Latn, oci_Latn, cat_Latn, fao_Latn, xho_Latn, spa_Latn, ron_Latn, amh_Ethi, ces_Latn, swe_Latn, nld_Latn, tat_Cyrl, kor_Hang, glg_Latn, fra_Latn, eus_Latn, ind_Latn, dan_Latn, tha_Thai, deu_Latn, tel_Telu, afr_Latn, pol_Latn, est_Latn, uig_Arab, ukr_Cyrl, uzn_Latn, heb_Hebr, kaz_Cyrl, nob_Latn, rus_Cyrl, vie_Latn, arb_Arab, zho_Hans, tuk_Latn, khk_Cyrl, jpn_Jpan, ell_Grek, isl_Latn, tam_Taml, slv_Latn, tur_Latn, mkd_Cyrl, tgl_Latn, gle_Latn"** as "Head" languages, and the remaining 135 languages (excluded English data) as "Long-tail" ones.

[3] `https://en.wikinews.org/wiki/MainPage`

[4] `https://en.wikibooks.org/wiki/Wikijunior`

[5] `https://en.wikivoyage.org/wiki/Main_Page`

Task category	Task	Train	Dev	Test	Lang.	Metric	Domain
Inference	AmericasNLI	392,702	222-743	738-750	10	Accuracy	Misc.
	XNLI	392,702	2,490	5,010	15	Accuracy	Misc.
Semantic Similarity	Multi-STS	550,152+5,749	10,000+1,500	250	7	Spearman	Misc.
	WMT21QETask1	7,000	1,000	1,000	7 (11)	Pearson	News
Sentence Retrieval	LAReQA	87,599	10,579	1,190	11	mAP@20	Wikipedia
	Mewsli-X	116,093	10,252	428-1,482	11 (50)	mAP@20	News
	BUCC	-	-	1,896-14,330	5	F1	Wiki/News
	Tatoeba	-	-	1,000	36 (122)	Accuracy	Misc.
Classification	XCOPA	33,410+400	100	500	11	Accuracy	Misc.
	MultiEURLEX	55,000	5,000	5,000	23	Accuracy	Legal
	MultiARC	200,000	5,000	5,000	6	MAE	Reviews
	PAWS-X	49,401	2,000	2,000	7	Accuracy	Wiki/Quora

Table 3: Overview of XRETE tasks. For tasks that have training and dev sets in other language, we only report the number of sentences in English sets. We report the number of test examples per languages.

B.2 Three measurements in Geometric Analysis

Invariance Measurement implies whether the semantic distributions of all languages are similar [Abend and Rappoport, 2017]. We adopt a Gaussian form $\mathcal{N}_l(\mu_l,\sigma_l^2)$ where $\mu_l=\frac{\sum_{\mathbf{x}\in l}\gamma^{(\mathbf{x})}}{3001}$ and $\sigma_l^2=\sum_{\mathbf{x}\in l}(\gamma^{(\mathbf{x})}-\mu_l)(\gamma^{(\mathbf{x})}-\mu_l)^T$, to approximate the semantic distribution of each language l. Further, we compute the mean averaged KL-divergence (KL-D for short) [Kullback and Leibler, 1951] among all language pairs as the overall Invariance score $\boldsymbol{\mathcal{I}_v}$ with L as the total number of languages:

$$\mathcal{I}_v = \frac{1}{L \times (L-1)} \sum_{l_1 \neq l_2} \frac{\mathbf{KL}(\mathcal{N}_{l_1}||\mathcal{N}_{l_2}) + \mathbf{KL}(\mathcal{N}_{l_2}||\mathcal{N}_{l_1})}{2}. \tag{1}$$

Canonical Form Measurement Previous works [Teller, 2000, Irwin et al., 2009] have demonstrated that a good multilingual space should distribute sentence representations based on their semantic similarities rather than language families. To measure this in quantity, we focus on Calinski-Harabasz Index (CH-I) [Caliński and Harabasz, 1974], which measures how similar an object is to its own cluster compared to other clusters. We group all semantically equivalent sentences in a cluster, which leads to 3001 clusters and each obsesses 204 sentences in 204 different languages. Assuming c_k and c are the centroid of cluster k and the whole dataset $\mathcal{S}$, respectively. The CH-I $\boldsymbol{\mathcal{C}_h}$ is defined as:

$$\mathcal{C}_h = \Big[204 \times \sum_{k=1}^{K} \|c_k - c\|^2\Big] / \Big[\sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \|s - c_k\|^2\Big]. \tag{2}$$

The higher the CH-I is, the better the semantically equivalent sentences are clustered.

Isotropy Measurement A high-dimensional embedding space often demonstrates poor isotropy, and deteriorates into a low-dimensional manifold that greatly limits the expressive ability of the embedding space. We adopt principal ratio (PR) [Mu and Viswanath, 2018] to measure isotropy. Let $\boldsymbol{\mathcal{E}}$ be the sentence representation matrix, $\boldsymbol{\mathcal{V}}$ be the set of the eigenvectors of $\boldsymbol{\mathcal{E}}$, the Isotropy $\boldsymbol{\mathcal{I}_{so}}$ is

$$\mathcal{I}_{so} = \min_{v \in \mathcal{V}} \sum_{e \in \mathcal{E}} \exp(v^\mathsf{T} e) / \max_{v \in \mathcal{V}} \sum_{e \in \mathcal{E}} \exp(v^\mathsf{T} e). \tag{3}$$

The closer $\boldsymbol{\mathcal{I}_{so}}$ is to 1, the more isotropic the representation space is.

C XRETE: Cross-lingual Representation Transfer Evaluation

XRETE consists of 12 tasks that fall into four different categories. In our "translate-train-all" setting, we individually fine-tune models with English training set and its translated training sets on each task. Then we report the performance of our fine-tuned model. We give an overview in Table 3 and describe the task details as follows.

XNLI The Cross-lingual Natural Language Inference corpus Conneau et al. [2018] tasks the systems with reading two sentences and determining whether one entails the other, contradicts it, or neither (neutral). A crowdsourcing-based procedure is used for collecting English examples, which are later translated into ten target languages for evaluation. Training data stays consistent with the English training data of MultiNLI Williams et al. [2018]. For evaluation, we concatenate two sentences as input and apply a new classification head to distinguish sentence relationships. We perform "translate-train-all" evaluation, where model is first fine-tuned on English training data and its translated data in other languages, then evaluated on test sets.

AmericasNLI (ANLI) The AmericasNLI Ebrahimi et al. [2022] is an extension of XNLI task to 10 Indigenous languages of the Americas. All of these languages are truly low-resource languages and serve as a good testbed for zero-shot cross-lingual transferability. As Spanish is more relative to the target languages, the Spanish version of XNLI subset is translated for evaluation. For training, both English and Spanish versions of MultiNLI training data are provided. We evaluate on ANLI following the same settings as in XNLI.

MultiSTS The Multilingual Semantic Textual Similarity dataset Cer et al. [2017], Reimers and Gurevych [2020] aims to assign a semantic similarity score for a pair of sentences. The MultiSTS dataset contains 7 cross-lingual sentence pairs and 3 monolingual pairs. Stanford NLI Bowman et al. [2015] and English STS Cer et al. [2017] are provided as training sets. We report the results after first fine-tuning on English training set using a Siamese network structure [Reimers and Gurevych, 2020]. Then we compute the cosine similarity between the sentence pairs and compute Spearman's rank correlation between the predicted score and gold score following Reimers and Gurevych [2020].

WMT21QETask1 (QE) The WMT21 Quality Estimation Task 1 Sentence-level Direct Assessment Specia et al. [2021] aims at testing the translation quality and this task has been applied to test the sensitivity of language models to semantic similarity Tiyajamorn et al. [2021]. The training and evaluation sets are collected from Wikipedia by translating sentences using state-of-the-art translation models to 6 languages and annotated by professional translators. In WMT21, 4 new language pairs with no training data are given to test zero-shot cross-lingual transferability. Our evaluation setting on QE is similar to that on MultiSTS, but we report Pearson's rank correlation [Kepler et al., 2019].

LAReQA The Language-Agnostic Retrieval Question Answering Roy et al. [2020] is a QA retrieval task where models are required to retrieve all relevant answers in different languages over a large multilingual pool. The dataset is constructed on XQuAD Artetxe et al. [2020] and a question is linked with answer sentences in different languages. The training set of SQuAD v1.1 Rajpurkar et al. [2016] is used to fine-tune the model to adapt to QA retrieval task. During evaluation, sentence embeddings are also obtained by a siamese network, and we retrieve the sentences with the highest cosine similarity as predictions.

Mewsli-X Mewsli (**M**ultilingual Entities in N**ews**, **li**nked) requires linking an entity mention to its entry in a language-agnostic knowledge base Botha et al. [2020]. Mewsli-X Ruder et al. [2021] features 15k mentions in 11 languages. For each mention, Mewsli-X offer entity description candidate pool containing 1M candidates across 50 languages. Fine-tuning is done on a predefined set of English-only mention-entity pairs from English Wikipedia hyperlinks. Our evaluation setting is identical to LAReQA.

BUCC The second and third shared task of the workshop on Building and Using Parallel Corpora Zweigenbaum et al. [2017], Pierre Zweigenbaum and Rapp [2018] aims to examine the ability of models to detect parallel sentence pairs in a pair of monolingual corpora. The dataset provides train and test splits in 5 languages. Following XTREME Hu et al. [2020], we directly evaluate on BUCC without fine-tuning and retrieve sentences with the highest cosine similarity.

Tatoeba The goal of the Tatoeba dataset Artetxe and Schwenk [2019] is to find the nearest neighbor for each sentence in the other language according to cosine similarity and compute the error rate. The dataset consists of up to 1,000 English-aligned sentence pairs covering 122 languages. Following XTREME Hu et al. [2020], we directly evaluate on Tatoeba without fine-tuning and retrieve sentences with the highest cosine similarity.

XCOPA In the Cross-lingual Choice of Plausible Alternatives dataset Ponti et al. [2020], each XCOPA instance corresponds to a premise and two alternatives. The task formulates as a binary

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	Avg.
XLM-R	88.6	84.5	86.7	84.6	85.2	84.7	82.0	82.5	82.6	82.4	80.6	83.1	80.3	77.3	77.2	**82.8**
INFOXLM	90.4	83.9	85.8	86.0	85.6	87.8	86.9	83.9	83.5	83.3	81.2	84.6	82.7	81.6	75.7	**84.2**
HICTL	90.6	86.8	88.2	87.4	87.0	87.4	85.0	83.9	83.3	84.8	83.1	85.7	82.8	79.7	80.9	**85.1**
ChatGPT	70.4	61.0	64.5	64.8	62.8	65.7	66.3	51.5	63.4	55.7	53.0	61.6	47.9	61.6	62.6	**60.9**
EMMA-X	**91.9**	**89.2**	**90.1**	**89.6**	**89.5**	**90.3**	**88.7**	**86.7**	**85.4**	**88.5**	**86.7**	**89.6**	**87.7**	**83.6**	**83.9**	**88.1**

Table 4: XNLI results (accuracy) for each language.

classification to predict the more plausible choice. The English COPA Gordon et al. [2012] training set and Social IQa Sap et al. [2019] training data are used for fine-tuning, while the validation and test sets of English COPA are translated and re-annotated into 11 languages for evaluation.

MultiEURLEX The MultiEURLEX dataset Chalkidis et al. [2021] is a legal topic classification task which comprises 65k European Union (EU) laws in 23 official EU languages. The dataset provides multi-granular labels per document. The dataset is split into training, development, and test subsets chronologically, resulting in 55k training documents for 7 languages, and 5k each for development and test subsets in all 23 languages.

MultiARC The Multilingual Amazon Reviews Corpus Keung et al. [2020] is a large-scale collection of Amazon reviews for multilingual text classification in 6 languages. Different languages are directly gathered from the marketplaces in different countries. The goal is to predict the reviewer's rating on the 5-star scale using the test of the review as input. The data is clearly split into training (200,000 reviews), development (5,000 reviews), and test sets (5,000 reviews) for each language.

PAWS-X The Cross-lingual Paraphrase Adversaries from Word Scrambling Yang et al. [2019b] dataset requires to identify whether two sentences are paraphrases. A subset of the evaluation pairs in English PAWS Zhang et al. [2019] are human-translated into 6 typologically distinct languages for evaluation, while the English PAWS training set is used for training.

D Baseline Methods

To fairly evaluate the performance of EMMA-X, we choose XLM-R Conneau and Lample [2019] and its several derivatives as our baselines, which contain: (1) XLM-R, which applies multilingual MLM tasks as pre-training objectives on CCNet-100 corpus; (2) HICTL Wei et al. [2021], which continues training on XLM-R using hierarchical contrastive learning; and (3) INFOXLM, which is initialized with XLM-R and trains with cross-lingual contrast, multilingual MLM and TLM. Also, we compare EMMA-X to strong sentence models: (1) S-BERT [Reimers and Gurevych, 2020], which adopts multilingual knowledge distillation to extend monolingual sentence representations to multilingual. We use the strongest baseline, **XLM-R ← SBERT-paraphrase**, proposed in the original paper as a baseline. (2) LaBSE [Feng et al., 2022], which systematically combines several best methods, including: masked language modeling, translation language modeling [Conneau and Lample, 2019], dual encoder translation ranking [Guo et al., 2018], and additive margin softmax [Yang et al., 2019a], to learn cross-lingual sentence representations. It filters 17B monolingual sentences and 6B translation pairs for sentence representation learning. We take the best model, LaBSE with Customized Vocab as our baseline. We further report the zero-shot results on Large Language Model (LLM), ChatGPT, which is trained on a wide variety of multilingual sentences and instruction tuning based on Reinforcement Learning with Human Feedback [Christiano et al., 2017, Ouyang et al., 2022].

E Prompts for ChatGPT

In this section, we show the input prompts of ChatGPT on each task in Table 7.

Model	aym	bzd	cni	gn	hch	nah	oto	quy	shp	tar	Avg.
XLM-R	49.01	50.61	41.72	58.34	42.46	54.63	35.57	59.29	51.62	41.54	**48.48**
INFOXLM	49.87	51.29	42.41	58.83	43.07	55.25	36.14	59.87	52.20	42.12	**49.10**
HICTL	49.65	51.22	42.36	58.82	43.09	55.13	36.04	59.61	52.17	42.08	**49.02**
ChatGPT	42.0	43.6	40.8	40.4	40.0	43.8	41.1	43.1	42.0	40.0	**41.7**
EMMA-X	**51.19**	**52.50**	**43.62**	**59.88**	**44.31**	**55.44**	**39.16**	**60.14**	**52.84**	**43.10**	**50.21**

Table 5: AmericasNLI (ANLI) results (top-1 accuracy) across different input languages.

Model	en-ar	en-de	en-tr	en-es	en-fr	en-it	en-nl	ar-ar	en-en	es-es	Avg.
XLM-R	50.2	63.7	45.8	59.6	68.0	63.4	69.6	87.7	82.5	68.5	**65.9**
INFOXLM	81.7	80.3	79.9	79.1	80.6	83.4	81.2	86.7	87.2	81.7	**82.2**
HICTL	80.4	81.8	78.3	80.6	81.2	80.9	79.3	88.4	86.1	79.6	**81.6**
EMMA-X	**86.6**	**85.0**	**87.1**	**84.4**	**85.2**	**89.4**	**88.3**	**90.9**	**92.0**	**84.5**	**87.3**

Table 6: MultiSTS results (Spearman) across different input languages.

F Results of each Language

We show the details for tasks and all languages in Tables 4 (XNLI), 5 (AmericasNLI), 6 (MultiSTS), 8 (QE), 9 (LAReQA), 10 (Mewsli-X), 11 (XCOPA), 12 (BUCC) and 13 (PAWS-X).

G Equations and Theoretical Analysis

G.1 Details of Equations

Details of Gaussian Form $\mathcal{N}_r$ In EMMA-X, GMM classifier is introduced to determine the semantic rank of sentence pairs. The posterior probability $P_{\mathcal{G}}(\cdot)$ of GMM classifier is already discussed in Eq. 5. We show the explicit calculation of Gaussian form $\mathcal{N}_r(\gamma^{(\mathbf{x}_i)}, \gamma^{(\mathbf{y}_k)})$ as:

$$\mathcal{N}_r(\gamma^{(\mathbf{x}_i)} - \gamma^{(\mathbf{y}_k)} | \mu_r, \sigma_r) = \frac{\pi_r}{(2\pi)^{(d/2)} |\mathrm{diag}(\sigma_r)|} \cdot e^{\left(-\frac{1}{2} \left[(\gamma^{(\mathbf{x}_i)} - \gamma^{(\mathbf{y}_k)}) - \mu_r \right]^T \mathrm{diag}(\sigma_r^{-2}) \left[(\gamma^{(\mathbf{x}_i)} - \gamma^{(\mathbf{y}_k)}) - \mu_r \right] \right)}, \tag{4}$$

where d is the dimension of hidden states of $\gamma^{(\mathbf{x}_i)}$ and $\gamma^{(\mathbf{y}_k)}$.

Details of contrastive learning The training objective of cross-lingual encoder in EMMA-X is the ranking InfoNCE loss. We show the explicit expansion of this loss (Eq. 7) as:

$$
\begin{aligned}
&\mathcal{L}_{\mathbf{CTL}}(\mathcal{X},\mathcal{Y};\Theta_{\mathcal{M}}) = -\mathbb{E}_{\mathbf{x}_i\sim\mathcal{X}}\Bigg[\\
&\underbrace{\log \frac{\sum_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=1}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_k)]}{\tau_1}}}{\sum_{\mathbf{y}_t\sim\mathcal{Y}_{c^*_{\mathcal{G}}=1}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_t)]}{\tau_1}} + \sum_{\mathbf{y}_t\sim\mathcal{Y}_{c^*_{\mathcal{G}}=2}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_t)]}{\tau_1}} + \ldots + \sum_{\mathbf{y}_t\sim\mathcal{Y},\mathcal{Y}_{c^*_{\mathcal{G}}=4}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_t)]}{\tau_1}}}}_{\ell_1} \\
&+ \underbrace{\log \frac{\sum_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=2}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_k)]}{\tau_2}}}{\sum_{\mathbf{y}_t\sim\mathcal{Y}_{c^*_{\mathcal{G}}=2}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_t)]}{\tau_2}} + \sum_{\mathbf{y}_t\sim\mathcal{Y}_{c^*_{\mathcal{G}}=3}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_t)]}{\tau_2}} + \sum_{\mathbf{y}_t\sim\mathcal{Y}_{c^*_{\mathcal{G}}=4}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_t)]}{\tau_2}}}}_{\ell_2} \\
&+ \underbrace{\log \frac{\sum_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=3}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_k)]}{\tau_3}}}{\sum_{\mathbf{y}_t\sim\mathcal{Y}_{c^*_{\mathcal{G}}=3}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_t)]}{\tau_3}} + \sum_{\mathbf{y}_t\sim\mathcal{Y}_{c^*_{\mathcal{G}}=4}} e^{\frac{s[\gamma(\mathbf{x}_i),\gamma(\mathbf{y}_t)]}{\tau_3}}}}_{\ell_3} \Bigg],
\end{aligned}
\tag{5}
$$

where τ_r represents the temperature term. As small temperature τ tends to be less tolerant to similar samples, and large τ tends to cluster similar samples together [Wang and Liu, 2021], we empirically set $\tau_1 < \tau_2 < \tau_3 < \tau_4$, which remains the same as Hoffmann et al. [2022].

G.2 Theoretical Analysis

In this section, we provide detailed proof for Eq. 14 and Eq. 15. Next, we prove the feasibility of our dual supervision. GMM classifier clusters sentence pairs in terms of Euclidean distance, while cross-lingual encoder minimizes the covariance of each semantic relation rank via cosine distance. Finally, we prove that these two metrics are actually equivalent to each other in the unit hypersphere of the embedding space.

Proof of Eq. 14. We provide the derivation of Eq. 14. With the assumption that $P(\mathbf{x}_i,\mathbf{y}_k|c^*_{\mathcal{G}} = r,\Theta) \sim \mathcal{N}_r\big(\mathbf{x}_i - \mathbf{y}_k|\tilde{\mu}_r,\tilde{\sigma}_r\big)$, we have,

$$
\begin{aligned}
\sum_{\mathbf{x}_i\in\mathcal{X}}\sum_{\mathbf{y}_k\in\mathcal{Y}}\sum_{r=1}^{N} Q(r)\log\frac{P(\mathbf{x}_i,\mathbf{y}_k,r|\Theta)}{Q(r)} &\approx \sum_{\mathbf{x}_i\in\mathcal{X}}\sum_{\mathbf{y}_k\in\mathcal{Y}}\sum_{r=1}^{N}\log P(\mathbf{x}_i,\mathbf{y}_k|c^*_{\mathcal{G}} = r,\Theta) \\
&= \sum_{\mathbf{x}_i\in\mathcal{X}}\sum_{\mathbf{y}_k\in\mathcal{Y}}\sum_{r=1}^{N}\Big(\log\big(\frac{1}{(2\pi)^{(d/2)}|\tilde{\sigma}_r|^{1/2}}\big) \\
&+ \frac{1}{2}\big[(\mathbf{x}_i-\mathbf{y}_k)-\tilde{\mu}_r\big]^T\tilde{\sigma}_r^{-1}\big[(\mathbf{x}_i-\mathbf{y}_k)-\tilde{\mu}_r\big]\Big) \\
&\geq \sum_{r=1}^{N}\big[\sum_{\mathbf{x}_i\in\mathcal{X}}\sum_{\mathbf{y}_k\in\mathcal{Y}}(\mathbf{x}_i-\mathbf{y}_k)\big]^2 - 2\tilde{\mu}_r\sum_{\mathbf{x}_i\in\mathcal{X}}\sum_{\mathbf{y}_k\in\mathcal{Y}}(\mathbf{x}_i-\mathbf{y}_k) + n\tilde{\mu}_r^2 \\
&= \sum_{r=1}^{N} n^2\tilde{\mu}_r^2 - n\tilde{\mu}_r^2 \\
&= n(n-1)\sum_{r=1}^{N}\tilde{\mu}_r^2,
\end{aligned}
\tag{6}
$$

with n denoting the number of sentence pairs in semantic rank r. Here, we ignore the impact of $\tilde{\sigma}_r$.

Proof of Eq. 15. As we apply dual supervision, data in the contrastive label space also follows the distribution $\mathcal{N}_r\big(\mathbf{x}_i - \mathbf{y}_k|\tilde{\mu}_r, \tilde{\sigma}_r\big)$. Hence, under mild assumptions, we can get:

$$
\begin{aligned}
\mathcal{L}^{+}_{\mathbf{CTL}}(\mathcal{X},\mathcal{Y};\Theta_{\mathcal{M}}) &= \mathbb{E}_{\mathbf{x}_i\sim\mathcal{X}} \sum_{r=1}^{N-1} \log \sum_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=r}} e^{s[\gamma^{(\mathbf{x}_i)},\gamma^{(\mathbf{y}_k)}]} \\
&= \sum_{\mathbf{x}_i\in\mathcal{X}} \sum_{\mathbf{y}_k\in\mathcal{Y}} \sum_{r=1}^{N-1} s(\mathbf{x}_i,\mathbf{y}_k) \\
&= \sum_{\mathbf{x}_i\in\mathcal{X}} \sum_{\mathbf{y}_k\in\mathcal{Y}} \sum_{r=1}^{N-1} \frac{(\mathbf{x}_i-\mathbf{y}_k)^2-2}{2} \\
&= n^2 \sum_{r=1}^{N-1} \tilde{\mu}_r^2.
\end{aligned}
\tag{7}
$$

Based on the definition of semantic ranks, we have $\tilde{\mu}_1 < \tilde{\mu}_2 < ... < \tilde{\mu}_N$. Empirically, the number of sentence pairs in each rank n is larger than the number of semantic ranks N. Hence, it can be derived that:

$$
\begin{aligned}
\mathcal{L}^{+}_{\mathbf{CTL}}(\mathcal{X},\mathcal{Y};\Theta_{\mathcal{M}}) &= n^2 \sum_{r=1}^{N-1} \tilde{\mu}_r^2 \\
&< n^2 \sum_{r=1}^{N-1} \tilde{\mu}_r^2 + n^2\tilde{\mu}_N^2 - n\sum_{r=1}^{N}\tilde{\mu}_r^2 \\
&= n(n-1)\sum_{r=1}^{N}\tilde{\mu}_r^2 \\
&\leq \sum_{\mathbf{x}_i\in\mathcal{X}} \sum_{\mathbf{y}_k\in\mathcal{Y}} \sum_{r=1}^{N} Q(r)\log\frac{P(\mathbf{x}_i,\mathbf{y}_k,r|\Theta)}{Q(r)}.
\end{aligned}
\tag{8}
$$

Therefore, we prove that minimizing the positive terms $\mathcal{L}^{+}_{\mathbf{CTL}}(\mathcal{X},\mathcal{Y};\Theta_{\mathcal{M}})$ in contrastive learning is equivalent to maximizing a lower bound of the likelihood in Eq. 12.

Feasibility of Dual Supervision According to the definition of semantic ranks, the approximated semantic rank $c^*_{\mathcal{G}}$ from GMM classifier should satisfy the following restriction,

$$
\mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=1}}||\gamma^{(\mathbf{x}_i)}-\gamma^{(\mathbf{y}_k)}|| < \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=2}}||\gamma^{(\mathbf{x}_i)}-\gamma^{(\mathbf{y}_k)}|| < ... < \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=N}}||\gamma^{(\mathbf{x}_i)}-\gamma^{(\mathbf{y}_k)}||. \tag{9}
$$

Similary, the approximated semantic rank $c^*_{\mathcal{M}}$ from cross-lingual encoder should satisfy the following restriction,

$$
\mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{M}}=1}} s[\gamma^{(\mathbf{x}_i)},\gamma^{(\mathbf{y}_k)}] > \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{M}}=2}} s[\gamma^{(\mathbf{x}_i)},\gamma^{(\mathbf{y}_k)}] > ... > \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{M}}=N}} s[\gamma^{(\mathbf{x}_i)},\gamma^{(\mathbf{y}_k)}]. \tag{10}
$$

Next, we prove that these two restrictions are interchangeable with each other in a unit hypersphere. For simplicity, we consider only two ranks, but extending the explanation to more ranks is trivial. As the Euclidean distance is always larger than 0, we have:

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=1}}||\gamma^{(\mathbf{x}_i)}-\gamma^{(\mathbf{y}_k)}|| < \mathbb{E}_{\mathbf{y}\sim\mathcal{Y}_{c^*_{\mathcal{G}}=2}}||\gamma^{(\mathbf{x}_i)}-\gamma^{(\mathbf{y}_k)}|| \\
&\qquad\Leftrightarrow \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=1}}(\gamma^{(\mathbf{x}_i)}-\gamma^{(\mathbf{y}_k)})^2 < \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=2}}(\gamma^{(\mathbf{x}_i)}-\gamma^{(\mathbf{y}_k)})^2 \\
&\qquad\Leftrightarrow \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=1}}(2-2\gamma^{(\mathbf{x}_i)}\gamma^{(\mathbf{y}_k)}) < \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=2}}(2-2\gamma^{(\mathbf{x}_i)}\gamma^{(\mathbf{y}_k)}) \\
&\qquad\Leftrightarrow \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=1}} s[\gamma^{(\mathbf{x}_i)},\gamma^{(\mathbf{y}_k)}] > \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{G}}=2}} s[\gamma^{(\mathbf{x}_i)},\gamma^{(\mathbf{y}_k)}] \\
&\qquad\Leftrightarrow \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{M}}=1}} s[\gamma^{(\mathbf{x}_i)},\gamma^{(\mathbf{y}_k)}] > \mathbb{E}_{\mathbf{y}_k\sim\mathcal{Y}_{c^*_{\mathcal{M}}=2}} s[\gamma^{(\mathbf{x}_i)},\gamma^{(\mathbf{y}_k)}].
\end{aligned}
\tag{11}
$$

From the above analyses, we can tell that the approximated semantic rank from one module can provide a reasonable supervision signal to guide the training of the other module. Hence, all sentence pairs will be uniformly distributed according to a unified ranking semantic similarity in the embedding space.

References

Omri Abend and Ari Rappoport. The state of the art in semantic representation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 77–89, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1008. URL https://aclanthology.org/P17-1008.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.1.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL https://aclanthology.org/Q19-1038.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421.

Jan A. Botha, Zifei Shan, and Daniel Gillick. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.630. URL https://aclanthology.org/2020.emnlp-main.630.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 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, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://www.aclweb.org/anthology/D15-1075.

T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101. URL https://www.tandfonline.com/doi/abs/10.1080/03610927408827101.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.559. URL https://aclanthology.org/2021.emnlp-main.559.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.280. URL https://aclanthology.org/2021.naacl-main.280.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://www.aclweb.org/anthology/D18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.435. URL https://aclanthology.org/2022.acl-long.435.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.480. URL https://aclanthology.org/2020.emnlp-main.480.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.62. URL https://aclanthology.org/2022.acl-long.62.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL https://aclanthology.org/S12-1052.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6317. URL `https://aclanthology.org/W18-6317`.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735. IEEE Computer Society, 2020.

David T Hoffmann, N Behrmann, J Gall, Thomas Brox, and M Noroozi. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. In *AAAI Conference on Artificial Intelligence*, 2022.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/hu20b.html`.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. 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 2485–2494, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1252. URL `https://aclanthology.org/D19-1252`.

Jeannie Y Irwin, Henk Harkema, Lee M Christensen, Titus Schleyer, Peter J Haug, and Wendy W Chapman. Methodology to develop and evaluate a semantic representation for nlp. In *AMIA Annual Symposium Proceedings*, volume 2009, page 271. American Medical Informatics Association, 2009.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-3020. URL `https://aclanthology.org/P19-3020`.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.369. URL `https://aclanthology.org/2020.emnlp-main.369`.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, San Diego, CA*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL `https://aclanthology.org/D18-2012`.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951. doi: 10.1214/aoms/1177729694. URL `https://doi.org/10.1214/aoms/1177729694`.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation. 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 3980–3994, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.308. URL `https://aclanthology.org/2021.acl-long.308`.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=HkuGJ3kCb`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=TG8KACxEON`.

Serge Sharoff Pierre Zweigenbaum and Reinhard Rapp. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In Reinhard Rapp, Pierre Zweigenbaum, and Serge Sharoff, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-07-8.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL `https://aclanthology.org/2020.emnlp-main.185`.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://aclanthology.org/D16-1264`.

Nils Reimers and Iryna Gurevych. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.365. URL `https://aclanthology.org/2020.emnlp-main.365`.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.477. URL `https://aclanthology.org/2020.emnlp-main.477`.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.802. URL `https://aclanthology.org/2021.emnlp-main.802`.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL `https://aclanthology.org/D19-1454`.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.507. URL https://aclanthology.org/2021.acl-long.507.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.71.

Virginia Teller. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Computational Linguistics*, 26(4):638–641, 12 2000. ISSN 0891-2017. doi: 10.1162/089120100750105975. URL https://doi.org/10.1162/089120100750105975.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.612. URL https://aclanthology.org/2021.emnlp-main.612.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30, NIPS 2017 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, June 2021.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. On learning universal representations across languages. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Uu1Nw-eeTxJ.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 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, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494.

Adina Williams, Nikita Nangia, and Samuel Bowman. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5370–5378. International Joint Conferences on Artificial Intelligence Organization, 7 2019a. doi: 10.24963/ijcai.2019/746. URL https://doi.org/10.24963/ijcai.2019/746.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. 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 3687–3692, Hong Kong, China, November 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1382. URL https://aclanthology.org/D19-1382.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1131. URL `https://aclanthology.org/N19-1131`.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 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, May 2016. European Language Resources Association (ELRA). URL `https://aclanthology.org/L16-1561`.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2512. URL `https://www.aclweb.org/anthology/W17-2512`.

Basic Prompt for XNLI/ANLI

Task Description: Read the following and determine the relationship between Hypothesis and Premise. Choose relation from "contradiction", "neutral", or "entailment".

Hypothesis: Yo... no puedo pensar por qué deberías hablarme así, dijo ella, con menos de lo que le había asegurado antes.

Premise: Ella era una buena amiga de él, por esto le dolía que le hablara así.

Basic Prompt for MultiSTS

Task Description: Read the following sentences and measure the real-valued meaning similarity between these two sentences. You can choose the meaning similarity score, ranging from 0 for no meaning overlap to 5 for meaning equivalence.

Sentence1: A person is on a baseball team.

Sentence2: Eine Person spielt in einem Team Basketball.

Basic Prompt for QE

Task Description: Read the Source sentence and its Translation, and estimate the quality of the Translation. You can rate the translation from 0-1 according to the perceived translation quality.

Source: În Franța a început stagnarea demografică de lungă durată, refacerea durând o generație.

Translation: In France, long-term demographic stagnation has started, restoring a generation.

Basic Prompt for XCOPA

Task Description: Read the Premise and determine which choice is the effect(or cause) of the Premise . Choose from "Choice1" or "Choice2".

Premise: Kuki kurukuna wasiman haykurqanku.

Choice1: Kuki kurukunaqa wasimanta chinkarqanku.

Choice2: Kuki kuruqa wasip kurkunta mikhurqanku.

Basic Prompt for MultiEURLEX

Task Description: Read the following sentences and determine the legal topic of the given sentence. Legal topic should choose from 'international organisations', 'social questions', 'production', 'technology and research', 'environment', 'energy', 'transport', 'law', 'finance', 'education and communications', 'trade', 'agriculture', 'forestry and fisheries', 'economics', 'agri-foodstuffs', 'EUROPEAN UNION', 'science', 'politics', 'international relations', 'industry', 'geography', 'business and competition', 'employment and working conditions'.

Sentence: NEUVOSTON ASETUS (EU) N:o 1390/2013, annettu 16 päivänä joulukuuta 2013, Euroopan unionin ja Komorien liiton kesken näiden välisessä kalastuskumppanuussopimuksessa määrätyjen kalastusmahdollisuuksien ja taloudellisen korvauksen vahvistamisesta hyväksytyn pöytäkirjan mukaisten kalastusmahdollisuuksien jakamisesta ...

Basic Prompt for MultiARC

Task Description: Read the following review and predict a 5-star scale rating (1 means the poorest experience and 5 represents excellent or outstanding performance) that can best match the review.

Review: no me llego el articulo me lo mando por correos normal sin seguimiento y nunca me llego tota un desastre

Basic Prompt for PAWS-X

Task Description: Read the following sentences and determine whether two sentences are paraphrases. Return yes or no.

Sentence1: La excepción fue entre fines de 2005 y 2009 cuando jugó en Suecia con Carlstad United BK, Serbia con FK Borac Čačak y el FC Terek Grozny de Rusia.

Sentence2: La excepción se dio entre fines del 2005 y 2009, cuando jugó con Suecia en el Carlstad United BK, Serbia con el FK Borac Čačak y el FC Terek Grozny de Rusia.

Table 7: Prompts of ChatGPT on each task.

Model	en-de	en-zh	et-en	ne-en	ro-en	ru-en	si-en	en-cs	en-ja	km-en	ps-en	Avg.
XLM-R	0.412	0.566	0.797	0.812	0.891	0.774	0.578	0.547	0.335	0.612	0.635	**0.632**
INFOXLM	0.517	0.534	0.775	0.834	0.890	0.788	0.581	0.564	0.325	0.635	0.616	**0.641**
HICTL	0.495	0.579	0.792	0.835	**0.904**	0.787	0.575	0.556	0.342	0.625	0.648	**0.649**
EMMA-X	**0.580**	**0.589**	**0.809**	**0.854**	0.897	**0.829**	**0.593**	**0.577**	**0.370**	**0.641**	**0.651**	**0.672**

Table 8: WMT21-QE-Task1 results (Pearson) across different input languages.

Model	ar	de	el	en	es	hi	ru	th	tr	vi	zh	Avg.
XLM-R	34.1	42.4	39.3	44.8	44.0	37.3	41.7	38.6	40.9	40.4	39.5	**40.3**
INFOXLM	39.7	52.6	39.2	55.1	53.4	36.8	51.0	28.5	41.1	48.9	47.3	**44.9**
HICTL	40.3	53.2	41.7	56.3	54.3	39.6	51.7	30.1	42.8	48.9	48.5	**46.1**
EMMA-X	**45.1**	**58.4**	**45.4**	**60.6**	**59.8**	**41.4**	**56.3**	**34.7**	**47.1**	**54.6**	**53.4**	**50.6**

Table 9: LAReQA results (mean average precision@20, mAP@20) across different input languages.

Model	ar	de	en	es	fa	ja	pl	ro	ta	tr	uk	Avg.
XLM-R	34.6	66.0	62.6	64.8	27.1	47.8	64.8	33.7	17.8	62.3	53.2	**48.6**
INFOXLM	40.8	71.6	66.3	68.7	48.7	61.0	66.7	39.2	42.0	64.6	58.1	**57.1**
HICTL	41.7	68.5	64.2	65.6	45.6	51.9	67.6	40.4	32.8	65.5	58.9	**54.8**
EMMA-X	**50.2**	**78.7**	**69.1**	**63.7**	47.9	**59.6**	**70.0**	**50.2**	**43.5**	**68.0**	**60.9**	**59.6**

Table 10: Mewsli-X results (mean average precision@20, mAP@20) across different input languages.

Model	et	ht	id	it	qu	sw	ta	th	tr	vi	zh	Avg.
XLM-R	73.8	67.4	77.8	72.2	52.3	70.9	72.1	**74.6**	73.4	73.2	75.7	**71.2**
INFOXLM	75.1	73.4	**78.3**	80.7	65.6	69.1	72.7	73.9	76.9	77.8	77.5	**74.6**
HICTL	75.9	73.1	77.8	**81.2**	65.5	73.8	72.6	73.2	76.1	75.4	78.0	**74.8**
ChatGPT	80.6	64.1	85.6	89.2	47.4	75.9	56.4	67.3	82.2	81.5	85.8	**74.2**
EMMA-X	**76.8**	**74.0**	77.6	79.8	**76.2**	**74.4**	**77.8**	74.2	**77.6**	**82.6**	**89.6**	**78.2**

Table 11: XCOPA results (accuracy) across different input languages.

Model	de	fr	ru	zh	Avg.
XLM-R	76.1	72.3	62.3	60.8	**67.9**
INFOXLM	81.3	78.2	76.0	74.2	**77.4**
HICTL	80.5	79.2	76.0	74.8	**77.6**
EMMA-X	**85.1**	**82.8**	**81.3**	**78.3**	**81.9**

Table 12: BUCC results (F1) across different languages.

Model	en	de	es	fr	ja	ko	zh	Avg.
XLM-R	95.7	92.2	92.7	92.5	84.7	85.9	87.1	**90.1**
INFOXLM	**97.7**	94.6	**95.2**	95.1	88.9	89.0	90.2	**93.0**
HICTL	97.4	94.2	95.0	94.2	89.1	89.5	90.2	**92.8**
ChatGPT	71.9	67.8	67.9	67.0	58.3	54.7	61.4	**64.2**
EMMA-X	97.3	**95.6**	94.7	**96.0**	**92.9**	**89.8**	**93.0**	**94.2**

Table 13: PAWS-X results (accuracy) for each language.