Analogy Training Multilingual Encoders*

Nicolas Garneau[†1], Mareike Hartmann[†2], Anders Sandholm[†3], Sebastian Ruder[†4], Ivan Vulić[†5] and Anders Søgaard[†2,3]

[1] Université Laval
[2] University of Copenhagen
[3] Google Research
[4] DeepMind
[5] University of Cambridge
nicolas.garneau@ift.ulaval.ca, {hartmann, soegaard}@di.ku.dk, sandholm@google.com, ruder@google.com, iv250@cam.ac.uk

Abstract

Language encoders encode words and phrases in ways that capture their local semantic relatedness, but are known to be globally inconsistent. Global inconsistency can seemingly be corrected for, in part, by leveraging signals from knowledge bases, but previous results are partial and limited to monolingual English encoders. We extract a large-scale multilingual, multi-word analogy dataset from Wikidata for diagnosing and correcting for global inconsistencies and implement a four-way Siamese BERT architecture for grounding multilingual BERT (mBERT) in Wikidata through analogy training. We show that analogy training not only improves the global consistency of mBERT, as well as the isomorphism of language-specific subspaces, but also leads to significant gains on downstream tasks such as bilingual dictionary induction and sentence retrieval.

Introduction

In NLP, there is a pressing need to build systems that bridge the digital language divide and serve all of the world's 7,000+ languages (Ruder, Vulić, and Søgaard 2019; Ponti et al. 2019). One research direction is to leverage similarities between languages for *cross-lingual transfer* (Snyder, Naseem, and Barzilay 2009; McDonald, Petrov, and Hall 2011; Täckström et al. 2013), e.g., through *general-purpose multilingual representations*, at the word level (Mikolov, Le, and Sutskever 2013; Faruqui and Dyer 2014; Artetxe, Labaka, and Agirre 2017) or at the sentence level and in context (Devlin et al. 2019; Lample and Conneau 2019).

Such pre-trained multilingual models have been shown to be surprisingly effective at cross-lingual transfer for some tasks (Pires, Schlinger, and Garrette 2019; Wu and Dredze 2019). Transfer is often simply a result of word-level alignment, however (Artetxe, Ruder, and Yogatama 2020)—and limited for more complex tasks and distant language families (Singh et al. 2019; Hu et al. 2020). Similar deficiencies have been observed for cross-lingual word embeddings (Gladkova, Drozd, and Matsuoka 2016; Vulić et al. 2019; Glavaš et al. 2019), where transfer has shown to be limited by the fact that word embedding spaces in different languages are often locally isomorphic, but not globally so (Søgaard, Ruder, and Vulić 2018; Nakashole and Flauger 2018; Schuster et al. 2019; Wu et al. 2020).

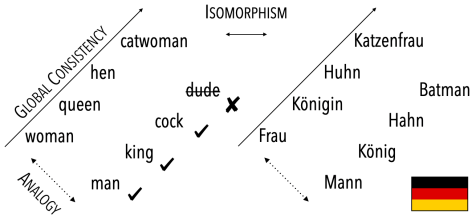


Figure 1: Encoders are globally inconsistent.

In this work, we hypothesize non-isomorphism at the global level is a result of global inconsistency (see Figure 1). Word embeddings often capture semantic similarities and analogies (Mikolov et al. 2013; Levy and Goldberg 2014), but only for short-range relations (Rogers, Drozd, and Li 2017). While the vector for *king* in Figure 1 may be predictable from the vectors of *woman*, *queen*, and *man*; and the vector for *cock* may be predictable from *woman*, *hen*, and *man*; the further we get away from *woman* and *man*, this effect degrades.[1] Encouraging spaces to be globally consistent and isomorphic is not straightforward (Zhang et al. 2019; Patra et al. 2019). Inspired by the observation that consistently encoding linguistic analogies entails isomorphism in the limit (Peng et al. 2020), we train multilingual encoders to encode analogies in order to encourage global consistency for cross-lingual transfer. Since existing analogy datasets are either monolingual or limited in size and the relations

*Code is available here: https://github.com/coastalcph/sentence-transformers-for-analogies

†Authors contributed equally to this work.

[1] http://bionlp-www.utu.fi/wv_demo/

they capture (Abdou, Kulmizev, and Ravishankar 2018), we present *WiQueen*[2], a large-scale analogy dataset across 11 languages based on publicly available Wikidata data. Using this dataset, we demonstrate that the embedding spaces of state-of-the-art monolingual and multilingual language encoders—similar to their word-level counterparts (Gladkova, Drozd, and Matsuoka 2016)—fail to capture long-distance relations. To address this global inconsistency, we propose a new four-way Siamese architecture to ground pre-trained language models in analogy relations. We show that this improves both analogy retrieval and the global consistency of the pre-trained embedding space. Finally, we also present downstream evaluations of our new, improved multilingual pretrained encoder.

Contributions We make publicly available a large-scale, multilingual multi-word analogy dataset for 11 languages, based on Wikidata. We present a new four-way Siamese architecture to ground the multilingual BERT model (Devlin et al. 2019) in this data; as well as a similar method for fastText word embeddings (Bojanowski et al. 2017). We evaluate the downstream impact of grounding mBERT on benchmark datasets for bilingual dictionary induction and sentence retrieval. We empirically validate that the grounding helps with global inconsistencies of pretrained language encoders and makes language-specific subspaces more isomorphic.

Quantifying the Global Consistency of Embedding Spaces

Multilingual representation learning relies on the ability to learn isomorphic representations for different languages (Mikolov et al. 2013; Søgaard, Ruder, and Vulić 2018). Monolingual representations cluster related words and phrases, but occasionally also exhibit global structure, e.g., the angle between verbs' present and past forms is near-constant (Levy and Goldberg 2014), and the same holds for words and their hyponyms. We refer to the degree to which a word embedding space exhibits this form of global structure as its global consistency: Static word embeddings and pretrained language encoders are globally consistent if the extent to which they reflect semantic relations is independent of scale; if the relations only hold locally, we call the models globally inconsistent.

We argue the global consistency of an encoder can be measured by the precision of its analogical reasoning; or, more precisely, the degree to which this precision drops when analogies span large distances in the embedding space. Analogical reasoning relies on the consistent encoding of semantic relations. If relations are encoded consistently, then, in the limit, the embedding space is isomorphic with other consistent embedding spaces (Peng et al. 2020). There have been controversies around analogies, however, which we briefly review, before we proceed.

[2]Available here: https://bit.ly/3aaKTzF

Controversies around Analogies (Levy and Goldberg 2014) showed that while word embeddings encode some linguistic relations in systematic ways, analogies based on other relations were not easily retrievable using simple vector offset. Gladkova, Drozd, and Matsuoka (2016) introduced the Bigger Analogy Test Set (BATS) dataset for English on which state-of-the-art word embeddings exhibited very low scores.[3] (Linzen 2016) further showed accuracy drops if the other elements of the analogy are not excluded as possible answer candidates—an observation later reiterated by Schluter (2018). (Rogers, Drozd, and Li 2017), inspired by observations of Levy and Goldberg (2014), showed that the accuracy of analogy retrieval decreases as the elements' distance in vector space increases. Our experiments demonstrate that representations learned by deep pre-trained models exhibit the same deficiency.

None of the mentioned critical studies, however, address the fundamental assumption that the global consistency or isomorphism that follows from analogical reasoning is a reasonable objective for word embeddings. They merely show that existing word embeddings may not encode analogical relations to the extent it was assumed before. Schluter (2018) conjectures that distributional information alone should not lead to analogical structure in the embedding space, but this is in apparent contrast to the observation in the cross-lingual word embedding literature that often, independently trained word vectors are near-isomorphic (Conneau et al. 2018a; Hartmann, Kementchedjhieva, and Søgaard 2018), as well as with observations made by Finley, Farmer, and Pakhomov (2017). Ethayarajh (2019) suggests that word embeddings may encode linguistic regularities as orthogonal transformations rather than translation vectors. While this assumption leads to better analogy retrieval, training language encoders to be consistent in encoding linguistic regularities this way is much more difficult than with simple vector offset.

Other studies indirectly motivate using analogies to improve word embeddings: (Drozd, Rogers, and Matsuoka 2016)—while critical of how analogies are used in practice—show that analogy retrieval can be improved by averaging the vector offset across similar analogies; this suggests that analogy training can be used to correct for idiosyncrasies and biases stemming from the underlying corpus sample. (Peng et al. 2020), more recently, derived the isomorphism of cross-lingual embedding spaces from the assumption that word embeddings exhibit analogical invariance. Their study, and the observations in Conneau et al. (2018a); Hartmann, Kementchedjhieva, and Søgaard (2018), also provide motivation for analogy training. We discuss the relationship between isomorphism and analogies in the Conclusion section.

WiQueen Since existing analogy datasets are either English-only or relatively small and limited in coverage, we

[3]BATS is much larger, more balanced, and more challenging than the Google Analogy dataset (Mikolov et al. 2013); BATS covers inflectional and derivational morphology, lexicographic, and encyclopedic semantics, where each relation is represented with 10 categories and each category contains 50 unique pairs.

introduce a new large-scale, multilingual analogy dataset. WiQueen[4] presents a set of 78,000 different analogies across 11 languages, linked to Wikidata[5] entities. The 11 languages were selected because they were already indexed in SLING,[6] which enables us to query Wikidata efficiently. The languages include: Danish, Dutch, English, Finnish, French, German, Italian, Polish, Portuguese, Spanish, and Swedish.[7]

We extracted the analogies from Wikidata as follows: We first identified all entities that are represented in Wikidata across all 11 languages. For these entities, we collected all pairs, (e_i, e_j) where there exists Wikidata property p such that $(e_i, e_j) \in p$. We recorded the types of the entities, $t(e_i)$ and $t(e_j)$. The type of e_i is $t(e_i)$ if $(e_i, t(e_i)) \in \textbf{instance\_of}$ in Wikidata. We grouped all pairs of triplets (e_i, e_j, p) (e_k, e_l, p) such that $t(e_i) = t(e_k)$ and $t(e_j) = t(e_l)$; e.g., the pair (London, England) is grouped with (Berlin, Germany) with other cities and countries in the **capital_of** relation. To generate analogies, we sample pairs of pairs from the groups. We realize analogies across all 11 languages to make our analogy datasets directly comparable across languages.

We present two versions of WiQueen. The first consists of the raw (11) files each containing all 78,000 analogies. The second is a subset of the WiQueen analogies where all four elements are predictable from the remaining elements. Using the terminology of Newman-Griffis, Lai, and Fosler-Lussier (2017), these are the analogies in which all relations are *informative*. Note this holds for many popular analogies, e.g., between pairs of countries and their capitals. The subset consists of 9,000 analogies and is created by only sampling from groups where all entities are unique, i.e., occur exactly once in the group's list of pairs. We provide each of these versions with standard training, validation and evaluation sections.

We augment the analogies with BigGraph (Lerer et al. 2019) distances, β, i.e. distances between graph embeddings of the entity nodes. We rely on these distances to quantify the global consistency, ρ, of pretrained language models, as the Pearson correlation between the cosine distances between entities (i.e. $\cos(\mathbf{v}_a, \mathbf{v}_3)$, where $\mathbf{v}_a = \mathbf{v}_1 - \mathbf{v}_2 + \mathbf{v}_4$) and the cosine distances between BigGraph representations of the analogy $\frac{1}{2}(\cos(\beta(e_1), \beta(e_2)) + \cos(\beta(e_3), \beta(e_4)))$. More precisely, if a pretrained language model is globally *inconsistent*, it will perform much worse on analogies that involve entities that are far apart, than on analogies that involve entities that are close. The distance of an analogy is *the average of the pairwise cosine distances of the two pairs of related entities' Wikidata graph nodes*.

Using Analogies to Quantify Global Consistency Rogers, Drozd, and Li (2017) have already empirically validated the tendency that analogy retrieval decreases with larger analogy distances in the (static) word embedding space. We now verify whether the same pattern also holds for pre-trained language models such as (m)BERT. We divide analogies into distance buckets according to their respective analogy distance (see the previous paragraph), and measure the precision (P@1) of nearest-neighbor-based analogy retrieval for each bucket. The first experiment conducted on the standard English BATS dataset (Gladkova, Drozd, and Matsuoka 2016) with a selection of English-pretrained LMs (as well as mBERT) verifies our intuition that these encoders, regardless of their training objective (cf., BERT and ELECTRA in Figure 2a) suffer from global inconsistency. Further, evaluating mBERT on the WiQueen data for all 11 languages indicates exactly the same pattern in all languages (Figure 2b): P@1 scores substantially decrease with larger analogy distances. These preliminary experiments inspire two crucial assumptions of this work: **(i)** there is ample room for improvement of global consistency in pretrained multilingual LMs such as mBERT, and **(ii)** instead of using word analogies only for intrinsic evaluation purposes, we should leverage the rich sources of information such as Wikidata to extract and inject analogical knowledge into pre-trained multilingual encoders for improved global consistency and, in consequence, improved task performance across languages. In what follows, we use the correlation between P@1 and analogy distances to quantify and monitor global consistency.

Analogy Training

This section provides algorithms for analogy training of static word embeddings and pretrained language models. We will use the analogies, i.e., instances of w_1 is to w_2 what w_3 is to w_4, to directly or indirectly minimize the following loss over the respective encodings/vectors $\mathbf{v}_{w_1}, \ldots, \mathbf{v}_{w_4}$:

$$\sum_{\langle w_1, w_2, w_3, w_4 \rangle} \cos(\mathbf{v}_{w_1} - \mathbf{v}_{w_2} + \mathbf{v}_{w_4}, \mathbf{v}_{w_3}) \qquad (1)$$

We present two algorithms for achieving this goal: for static word embeddings, e.g., fastText (Bojanowski et al. 2017), and for pre-trained language models, e.g., BERT (Devlin et al. 2019).

Analogy Training of Static Word Embeddings Given an analogy $\langle w_1, w_2, w_3, w_4 \rangle$, we would like to encourage the model to retrieve the correct target word w_3 given w_1, w_2 and w_4 based on simple vector offset. As analogies will not be available for all words in a model's vocabulary, we also need to propagate the analogical knowledge globally to the entire vector space including words that are not present in the input analogy set.

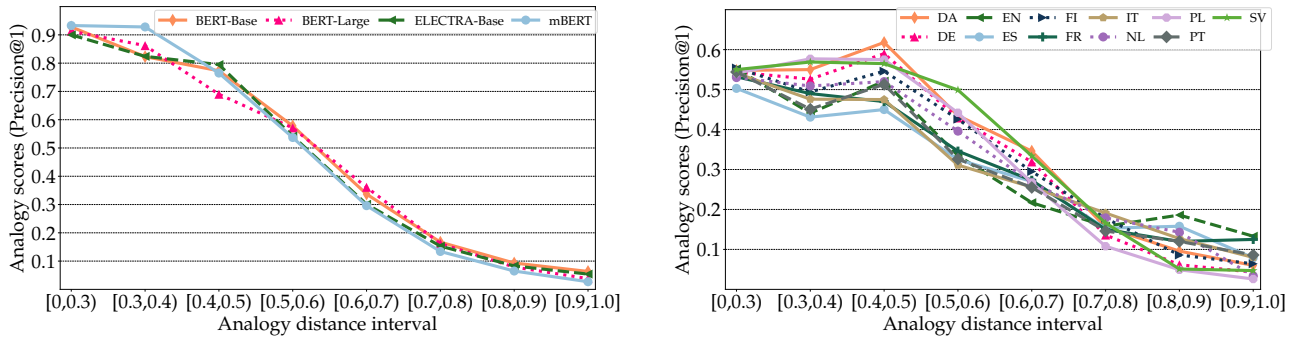
As most of our analogies are composed of multi-word expressions (MWEs), we use the mean of the fastText embeddings as the representation of an entity, i.e., $\mathbf{v}_w = \frac{1}{|w|} \sum_{i=1}^{|w|} \mathbf{v}_i$, where $|w|$ is the number of words that compose entity w, and $\mathbf{v}_i$ is the fastText's embedding of the i-th word in the entity w. We ignore out-of-vocabulary tokens, and analogies with no in-vocabulary tokens.

[4]The name is a reference to Wikidata, as well as the perhaps most famous analogy from the initial study of analogical reasoning with word embeddings (Mikolov et al. 2013).

[5]https://www.wikidata.org/

[6]https://github.com/ringgaard/sling

[7]We leave out Norwegian, for which entities overlap very little with the other languages.



(a) English BATS dataset

(b) WiQueen (11 languages)

Figure 2: Diagnostic experiments in analogy retrieval demonstrating that pre-trained language models are globally inconsistent. P@1 scores are reported over different analogy distance intervals/buckets. **(a)** P@1 on the English BATS dataset (Gladkova, Drozd, and Matsuoka 2016) with several standard encoders pre-trained on English data such as BERT-Base/Large (Devlin et al. 2019), and ELECTRA-Base (Clark et al. 2020), as well as mBERT; **(b)** P@1 on the 11 language-specific subsets of the WiQueen data using mBERT as the language encoder.

Let $\mathcal{B}$ be a mini-batch of analogies and their corresponding fastText averaged embeddings. We compute $\mathbf{v}_a = \mathbf{v}_1 - \mathbf{v}_2 + \mathbf{v}_4$, yielding a batch of k pairs, $\mathcal{B} = [(\mathbf{v}_a^1, \mathbf{v}_3^1), \ldots, (\mathbf{v}_a^k, \mathbf{v}_3^k)]$. We then draw a set of negative examples $\mathcal{N} = [(\mathbf{t}_a^1, \mathbf{t}_3^1), \ldots, (\mathbf{t}_a^k, \mathbf{t}_3^k)]$ where $\mathbf{t}_a^i$ is the nearest neighbor to $\mathbf{v}_a^i$, and $\mathbf{t}_3^i$ the nearest neighbor of $\mathbf{v}_3^i$.

Our method is different from the interactive method proposed by Yuan et al. (2020). We now encourage the model to bring the analogical pairs closer together in the embedding space compared to the negative examples. For this, we follow the attract part of the Attract-Repel (AR) algorithm (Mrkšić et al. 2017; Vulić et al. 2018) to perform analogy training. In the attract step, we minimize the loss $\mathcal{A}(\mathcal{B}, \mathcal{N})$:

$$\sum_{i=1}^{k} \left(\tau\left(\delta + \mathbf{v}_a^i \mathbf{t}_a^i - \mathbf{v}_a^i \mathbf{v}_3^i\right) + \tau\left(\delta + \mathbf{v}_3^i \mathbf{t}_3^i - \mathbf{v}_a^i \mathbf{v}_3^i\right)\right) \quad (2)$$

where $\tau(z) = \max(0, z)$ is the standard rectifier function (Nair and Hinton 2010) and δ is the margin that determines how much closer these vectors should be to each other compared to their respective negative examples. We add a regularization term to preserve the semantic information in the original distributional vector space:

$$\mathcal{R}(\mathcal{E}_\mathcal{B}) = \sum_{\mathbf{v}_i \in \mathcal{E}_\mathcal{B}} \lambda ||\hat{\mathbf{v}}_i - \mathbf{v}_i||_2 \quad (3)$$

where $\mathcal{E}_\mathcal{B}$ the set of all entity vectors present in a mini-batch, λ is the ℓ_2-regularisation constant, and $\hat{\mathbf{v}}_i$ denotes the original distributional word representation of entity w_i. The final cost function is then the sum of both terms: $\ell(\mathcal{B}, \mathcal{N}, \mathcal{E}_\mathcal{B}) = \mathcal{A}(\mathcal{B}, \mathcal{N}) + \mathcal{R}(\mathcal{E}_\mathcal{B})$.

However, the Attract-Repel algorithm fine-tunes only for the subspace of vectors of words present in external data (i.e., input analogies)—the subspace $\mathbf{V}_{seen}$. In order to propagate the analogical signal to the entire vector space, we learn a (global) mapping function (i.e., the so-called *post-specialization* mapping, see the work of Vulić et al. (2018) for further details) between the initial input vectors (i.e., $\hat{\mathbf{v}}_i$-s) and their refined "analogy-specialized" variants obtained after applying the AR procedure (i.e., $\mathbf{v}_i$-s). The mapping is realized as a deep feed-forward network similar to the one of Vulić et al. (2018); Zhang et al. (2020): we learn the mapping based on all analogy pairs in $\mathbf{V}_{seen}$ and apply it to all other vectors for words unseen in the analogy set, i.e., the subspace $\mathbf{V}_{unseen}$.

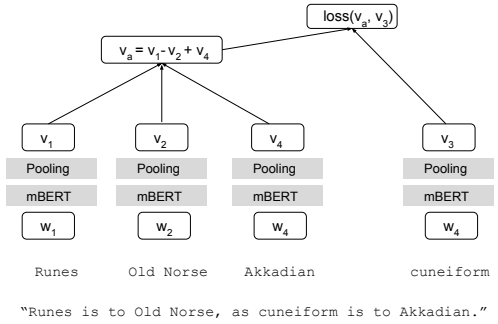


Figure 3: Siamese BERT for analogy training. Following the Sentence BERT architecture, fixed size embeddings for sequences of words (in our case single tokens or MWEs that represent an entity of the analogy) are computed by mean pooling the subtoken representations computed by the encoder.

Analogy Training using Siamese BERT In order to fine-tune a pre-trained language model such as BERT on the analogy retrieval task, we use a Siamese network architecture[8], which is shown in Figure 3. We embed the four en-

[8]The Siamese BERT network is similar to two-headed networks for semantic similarity fine-tuning, e.g., (Reimers and Gurevych 2019; Humeau et al. 2020; Henderson et al. 2019).

tities $\langle w_1, w_2, w_3, w_4 \rangle$ of the analogy using a Siamese network with four copies of BERT. Each copy consists of a pre-trained BERT body and a mean pooling layer at the output and produces $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, and $\mathbf{v}_4$ respectively. From the output of the pooling layers, we compute $\mathbf{v}_a = \mathbf{v}_1 - \mathbf{v}_2 + \mathbf{v}_4$ and minimize the distance between $\mathbf{v}_a$ and $\mathbf{v}_3$.

Different from the static word embeddings procedure, we experiment with two different objectives: minimizing the MSE loss $||\mathbf{v}_a - \mathbf{v}_3||^2$, or using a contrastive loss, computed as follows:

$$\max(||\mathbf{v}_a - \mathbf{v}_3|| - ||\mathbf{v}_a - \mathbf{v}_x|| + \epsilon, 0) \quad (4)$$

Here, $\mathbf{v}_3$ corresponds to the correct entity fitting in the analogy, whereas $\mathbf{v}_x$ is the embedding of an entity that does not fit into the analogy. This incorrect entity w_x is determined on-the-fly as the hardest negative within the batch, which has the smallest distance to $\mathbf{v}_a$. This loss enforces the correct entity to be closer to $\mathbf{v}_a$ than the incorrect entity. In our experiments, we find that a post-specialization equivalent term is not necessary for pre-trained models as all of the model's parameters are updated and are thus encouraged to capture global consistency during analogy training. When fine-tuning multilingual BERT, we train on the analogies across *all* languages simultaneously.

In our experiments, we also evaluate the effect of using aliases[9], i.e., alternative labels of the entities listed in Wikidata, as well as descriptions[10], to augment the entities with more context. Take for example the following analogy;

Hefei is to **Anhui**, as **Guiyang** is to **Guizhou**.

Without any contextual information, it may be hard for a model to reason about these particular entities. With aliases and descriptions, we can augment the above analogy as follows:

Hefei Luzhou, Hofei, capital of Anhui province, China is to **Anhui** province of China, as **Guiyang** capital of Guizhou province, China is to **Guizhou** province of China.

We use the special symbol `[SEP]` to concatenate aliases and the description. To prevent leakage, as with the entity **Guiyang**, we mask occurrences of w_3 in the other entities (w_1, w_2 and w_4), as well as the occurrence of any other entities in w_3's aliases and descriptions. This leads to masked, augmented analogies consisting of entities such as the following for **Hefei** and **Guiyang**:

Hefei `[SEP]` Luzhou, Hofei `[SEP]`
list of aliases
capital of Anhui province, China
description
Guiyang `[SEP]` capital of `[UNK]` province, China
description

We refer to the encoder fine-tuned on the augmented WiQueen as **WiQueen**$^+$. **WiQueen** has on average 60K tokens and a vocabulary size of 15K types, while **WiQueen**$^+$ has 250K and 25K, respectively. We can find 143 different types of analogies (e.g. `owner of`) and 534 types of entities (e.g. `association football club` and `stadium`) with 15,731 different instances (e.g. `Sunderland Football Club` and `Stadium of Light`).

[9]https://www.wikidata.org/wiki/Help:Aliases

[10]https://www.wikidata.org/wiki/Help:Description

Experiments

Intrinsic Evaluation

We first present an intrinsic evaluation in the analogy retrieval task of both baseline models and the proposed models after analogy training. We present results for static word embeddings and pre-trained language encoders before and after fine-tuning on the analogy data in Table 1. We report both P@1 and global consistency scores.

Static word embeddings For static cross-lingual word embeddings, we use the aligned, pre-trained fastText word vectors.[11] In the first and third pair of columns of Table 1, we evaluate the word vectors as-is on WiQueen. Analogy training of the multilingual fastText model, unsurprisingly, improves the precision of analogy retrieval with this model: P@1 improves by 0.025. The global consistency also improves a little by analogy training, but the effect is not as strong as with pre-trained language models.

Pre-trained language models With mBERT-WiQueen variants, we see similar improvements from analogy training:[12] a 0.11 (mBERT-WiQueen) to almost 0.15 (mBERT-WiQueen$^+$) improvement in P@1, but over a much stronger baseline than fastText. Interestingly, we also see a consistent and very significant effect on global consistency with both variants, with larger consistency using the augmented mBERT-WiQueen$^+$ variant. Across all language pairs, mBERT learned newest analogies in the spatial (e.g. `Tripoli` is to `Tripolitania` what `Thessaloniki` is to `Macedonia`) and temporal (e.g. `Cryptic Writings` (1997), album of Megadeth, follows `Youthanasia` (1994) as `Automatic for the People` (1992), album of R.E.M, follows `Out of Time` (1991)) types of analogies. mBERT also learned new types of analogies e.g. "occupant" in the context of sports teams; `Tampa Bay Lightning` is to `Amalie Arena` what `Minnesota Wild` is to `Xcel Energy Center`. In the Appendix, we also present results for XLM-R (Conneau et al. 2020), another multilingual pretrained language encoder. These results are lower and less consistent, both before and after analogy training, so we focus on mBERT here.

Extrinsic Evaluation

The intrinsic evaluation shows that analogy training improves global consistency across languages. Globally consistent encoders should enable more precise transfer across languages, and hence should provide improvements for cross-lingual NLP tasks. We evaluate analogy training on two downstream tasks that rely on the global geometry of

[11]https://fasttext.cc/docs/en/aligned-vectors.html

[12]The contrastive loss outperforms MSE loss for analogy training of mBERT, and we henceforth report results based on the contrastive loss.

Language	Baselines				Analogy Training					
	Fasttext		mBERT		Fasttext		mBERT-WiQueen		mBERT-WiQueen$^+$	
	P@1	ρ	P@1	ρ	P@1	ρ	P@1	ρ	P@1	ρ
Danish	0.1511	0.3001	0.2835	0.3221	0.1688	0.2909	0.3863	0.3010	**0.3935**	**0.2461**
German	0.0997	0.3604	0.2658	0.3548	0.1104	0.3702	0.3894	0.3257	**0.4538**	**0.2868**
English	0.1255	0.2854	0.2897	0.3107	0.1513	0.2550	0.4091	0.2960	**0.4787**	0.2821
Spanish	0.0899	0.3383	0.2596	0.3441	0.1194	0.3573	0.3832	0.3198	**0.3936**	**0.3012**
Finnish	0.1258	0.3908	0.2679	0.3535	0.1682	0.3731	0.3728	0.3192	**0.4019**	**0.2703**
French	0.0943	0.3659	0.2617	0.3545	0.1146	0.3459	0.3707	0.3375	**0.4195**	**0.2991**
Italian	0.0731	0.3979	0.2773	0.3711	0.0949	0.3883	0.3821	**0.3338**	**0.4372**	0.3722
Dutch	0.1291	0.3520	0.2669	0.3443	0.1497	0.3384	0.3811	**0.3202**	**0.4424**	0.3609
Polish	0.1165	0.3397	0.2648	0.3656	0.1456	0.3287	0.3853	0.3468	**0.3894**	**0.2718**
Portuguese	0.0898	0.3614	0.2523	0.3536	0.1072	0.3640	0.3697	0.3409	**0.3718**	**0.2653**
Swedish	0.1071	0.3449	0.2856	0.3378	0.1415	0.3270	0.3832	0.3128	**0.4071**	**0.2672**
Averages	0.1093	0.3488	0.2704	0.3435	0.1338	0.3399	0.3830	0.3231	**0.4171**	**0.2930**

Table 1: Evaluation of fasttext and mBERT embeddings on the WiQueen dataset. P@1 is the precision of analogical retrieval (Gladkova, Drozd, and Matsuoka 2016). ρ is a measure of global consistency as defined previously. Observations: **(a)** Analogy training, as expected, improves performance on analogy retrieval in 11/11 cases (ΔP@1.$\uparrow$). **(b)** Analogy training improves global consistency in 10/11 cases ($\Delta\rho\downarrow$). **(c)** Cross-lingual variation is limited.

		Tatoeba			BUCC2018		
	Language	**mBERT**	**mBERT-WiQueen**	**mBERT-WiQueen$^+$**	**mBERT**	**mBERT-WiQueen**	**mBERT-WiQueen$^+$**
SLING	Dutch	0.6370	0.6570	**0.6640**	–	–	–
	German	0.7540	0.7420	**0.7550**	0.6326	**0.6412**	0.6398
	French	0.6430	**0.6740**	0.6150	0.6246	**0.6435**	0.6388
	Finnish	0.3900	0.3820	**0.3910**	–	–	–
	Italian	0.5730	0.6070	**0.6170**	–	–	–
	Portuguese	0.6840	0.6910	**0.6980**	–	–	–
	Spanish	0.6410	**0.6720**	0.6470	–	–	–
Average SLING		0.6174	**0.6321**	0.6267	0.6286	**0.6423**	0.6393
Average		0.3753	0.3844	**0.3853**	0.5780	0.5991	**0.6027**

Table 2: SENTENCE RETRIEVAL results. We follow Hu et al. (2020) in reporting accuracy for Tatoeba (Artetxe and Schwenk 2019) and F_1 for BUCC2018 (Zweigenbaum, Sharoff, and Rapp 2017). Averages are for *all* languages in the benchmarks, including **languages that are not in SLING**, e.g., Chinese, Russian, etc. We see significant improvements from WiQueen training on these languages, too.

embedding spaces. We focus on tasks where embeddings can be evaluated directly in order to control for task fine-tuning as a source of variation in performance. In the Appendix, we, in addition, report results for XNLI (Conneau et al. 2018b).

Bilingual Dictionary Induction (BDI) is the task of inducing word-level translations with no or limited supervision. This can be done by learning a linear alignment between language-specific word embedding spaces. Pretrained multilingual language encoders have been shown to be effective for this task (Wu and Dredze 2019). We evaluate the pretrained encoders on their ability to induce translation pairs in the standard MUSE dictionaries (Conneau et al. 2018a).[13] Following (Conneau et al. 2018a), we encode all dictionary entries and 200,000 candidate words in the target language as the output of the encoders' pooling layer, and induce translations by querying nearest neighbors across languages using cosine similarity. Results, with English as the source language in all experiments, are shown in Table 3. While word-level translations can rarely be accurately retrieved in the embedding space of mBERT, analogy training substantially increases the retrieval precision across languages. Interestingly, training *without* aliases and descriptions is best, possibly because of context being unavailable at test time

[13] https://github.com/facebookresearch/MUSE. See (Kementchedjhieva, Hartmann, and Søgaard 2019) for a discussion of biases in the MUSE benchmark.

	mBERT	WiQueen	WiQueen$^+$
German	0.1161	**0.2578**	0.2166
French	0.1895	**0.3947**	0.2974
Italian	0.1661	**0.3287**	0.2367
Spanish	0.1781	**0.4262**	0.3099
Danish	0.1062	**0.2062**	0.1690
Finnish	0.0704	**0.1048**	0.1031
Dutch	0.1022	**0.2260**	0.1889
Polish	0.1044	**0.2216**	0.1689
Portuguese	0.1465	**0.3187**	0.2252
Swedish	0.0963	**0.1891**	0.1637

Table 3: BILINGUAL DICTIONARY INDUCTION results of mBERT, mBERT fine-tuned on WiQueen and WiQueen$^+$. We report P@10 scores of a nearest neighbor search with cosine similarity in the original embedding spaces. This is not the optimal approach to bilingual dictionary induction (Wu and Dredze 2019), but it directly evaluates the isomorphism of our encoders.

in BDI. The global structure of the language-specific subspaces nevertheless improves, i.e., the subspaces become more isomorphic. This is validated by applying two standard measures of isomorphism from prior work. We observe reductions in average Gromov-Hausdorff distance (Patra et al. 2019) across languages – from 0.66 in mBERT to 0.48 for mBERT-WiQueen and 0.46 for mBERT-WiQueen$^+$ – and in isospectrality (Søgaard, Ruder, and Vulić 2018) (from 82.6 to 44.3 and 34.2). See Appendix for full results.

Sentence Retrieval We use two standard sentence retrieval tasks, Tatoeba (Artetxe and Schwenk 2019) and BUCC2018 (Zweigenbaum, Sharoff, and Rapp 2017), for evaluating the downstream performance of multilingual analogy training. For Tatoeba, which consists of up to 1,000 English-aligned sentence pairs across 36 languages, we follow Hu et al. (2020) and query the nearest neighbour of the input sentence in the target sentences using cosine similarity and calculate the error rate. For BUCC2018—which covers only five languages (de, en, fr, ru, and zh)—we also use cosine similarity, but report F_1, again following Hu et al. (2020). Our mBERT baseline results are comparable to those of Hu et al. (2020). We observe improvements due to the analogy training for most languages and better average performance even if we include languages for which the model was not trained with analogy data.

Discussion and Conclusion

(Peng et al. 2020) try to derive the isomorphism of cross-lingual embedding spaces from the assumption that they exhibit analogical invariance. They present a proof that the linearity of cross-lingual mappings of embedding spaces depends on the preservation of analogical information encoded in monolingual vector spaces. This also follows from the definition of isomorphisms $\mathbf{T}$ of vector spaces, i.e., $\mathbf{T}(v + w) = \mathbf{T}(v) + \mathbf{w}$ and $\mathbf{T}(cv) = c\mathbf{T}(v)$. Analogy training should therefore lead to better bilingual dictionary induction results using nearest neighbor search between language-specific embedding spaces (Conneau et al. 2018a); this is confirmed by our results in §4.2.

Nakashole and Flauger (2018) claim isomorphism holds between geometrically-local regions of cross-lingual word embedding spaces rather than between the entire spaces. This would mean that only local analogies were invariant across language-specific embedding spaces. Our results indicate this tendency holds, and that analogy training can be used to correct for this deficiency in multilingual encoders. Similar assumptions have motivated seed extraction methods for unsupervised alignment of monolingual word embedding spaces (Aldarmaki, Mohan, and Diab 2018; Artetxe, Labaka, and Agirre 2018).

Other attempts to encourage isomorphism have been proposed, but, to the best of our knowledge, only for static word embeddings: Zhang et al. (2019) use iterative normalization to encourage isomorphism. Patra et al. (2019) use a mixture of explicit supervision and distributional information. Neither of the two algorithms is applicable to pretrained language encoders with dynamic, open-ended vocabularies.

We presented a novel, large-scale multilingual analogy dataset, WiQueen, covering 11 languages and a wide range of semantic relations, as well as algorithms for analogy training for pretrained language encoders and static word embeddings. We used the analogies to diagnose the global inconsistency of multilingual encoders. We evaluated our learning algorithms across intrinsic and extrinsic benchmarks and showed that analogy training improves the global consistency of multilingual encoders and leads to better performance in tasks that require globally consistent representations, such as bilingual dictionary induction and sentence retrieval.

References

Abdou, M.; Kulmizev, A.; and Ravishankar, V. 2018. MGAD: Multilingual Generation of Analogy Datasets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1320.

Aldarmaki, H.; Mohan, M.; and Diab, M. 2018. Unsupervised Word Mapping Using Structural Similarities in Monolingual Embeddings. *Transactions of the Association for Computational Linguistics* 6: 185–196. doi:10.1162/tacl_a_00014. URL https://www.aclweb.org/anthology/Q18-1014.

Artetxe, M.; Labaka, G.; and Agirre, E. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of ACL 2017*, 451–462.

Artetxe, M.; Labaka, G.; and Agirre, E. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL 2018*, 789–798. doi:10.18653/v1/P18-1073. URL https://www.aclweb.org/anthology/P18-1073.

Artetxe, M.; Ruder, S.; and Yogatama, D. 2020. On the Cross-lingual Transferability of Monolingual Representa-

tions. In *Proceedings of ACL 2020*. URL http://arxiv.org/abs/1910.11856.

Artetxe, M.; and Schwenk, H. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics* 7: 597–610. URL https://transacl.org/ojs/index.php/tacl/article/view/1742.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5: 135–146. doi:10.1162/tacl_a_00051. URL https://www.aclweb.org/anthology/Q17-1010.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *Proceedings of ICLR 2020*. URL https://openreview.net/forum?id=r1xMH1BtvB.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL 2020*.

Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2018a. Word Translation Without Parallel Data. In *Proceedings of ICLR 2018*.

Conneau, A.; Rinott, R.; Lample, G.; Williams, A.; Bowman, S.; Schwenk, H.; and Stoyanov, V. 2018b. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of EMNLP*, 2475–2485. doi:10.18653/v1/D18-1269. URL https://www.aclweb.org/anthology/D18-1269.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*, 4171–4186.

Drozd, A.; Rogers, A.; and Matsuoka, S. 2016. Word Embeddings, Analogies, and Machine Learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016*, 3519–3530.

Ethayarajh, K. 2019. Rotate King to get Queen: Word Relationships as Orthogonal Transformations in Embedding Space. In *Proceedings of EMNLP-IJCNLP 2019*, 3503–3508.

Faruqui, M.; and Dyer, C. 2014. Improving Vector Space Word Representations Using Multilingual Correlation. In *Proceedings of EACL 2014*, 462–471.

Finley, G.; Farmer, S.; and Pakhomov, S. 2017. What Analogies Reveal about Word Vectors and their Compositionality. In *Proceedings of StarSEM 2017*, 1–11.

Gladkova, A.; Drozd, A.; and Matsuoka, S. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop 2016*, 8–15. doi:10.18653/v1/N16-2002. URL https://www.aclweb.org/anthology/N16-2002.

Glavaš, G.; Litschko, R.; Ruder, S.; and Vulić, I. 2019. How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions. In *Proceedings of ACL 2019*, 710–721. URL https://arxiv.org/abs/1902.00508.

Hartmann, M.; Kementchedjhieva, Y.; and Søgaard, A. 2018. Why is unsupervised alignment of English embeddings from different algorithms so hard? In *Proceedings of EMNLP 2018*, 582–586. doi:10.18653/v1/D18-1056. URL https://www.aclweb.org/anthology/D18-1056.

Henderson, M.; Casanueva, I.; Mrkšić, N.; Su, P.-H.; Vulić, I.; et al. 2019. ConveRT: Efficient and Accurate Conversational Representations from Transformers. *arXiv preprint arXiv:1911.03688* URL https://arxiv.org/abs/1911.03688.

Hu, J.; Ruder, S.; Siddhant, A.; Neubig, G.; Firat, O.; and Johnson, M. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. *arXiv preprint arXiv:2003.11080* .

Humeau, S.; Shuster, K.; Lachaux, M.; and Weston, J. 2020. Poly-encoders: Transformer Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *Proceedings of ICLR 2020*. URL http://arxiv.org/abs/1905.01969.

Kementchedjhieva, Y.; Hartmann, M.; and Søgaard, A. 2019. Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary Induction. In *Proceedings EMNLP-IJCNLP 2019*, 3336–3341. doi:10.18653/v1/D19-1328. URL https://www.aclweb.org/anthology/D19-1328.

Lample, G.; and Conneau, A. 2019. Cross-lingual Language Model Pretraining. In *Proceedings of NeurIPS 2019*, 7057–7067.

Lerer, A.; Wu, L.; Shen, J.; Lacroix, T.; Wehrstedt, L.; Bose, A.; and Peysakhovich, A. 2019. PyTorch-BigGraph: A Large-scale Graph Embedding System. In *Proceedings of the 2nd SysML Conference*.

Levy, O.; and Goldberg, Y. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of CoNLL 2014*, 171–180. ISBN 9781941643020. URL http://anthology.aclweb.org/W/W14/W14-16.pdf{\#}page=181.

Linzen, T. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, 13–18.

McDonald, R.; Petrov, S.; and Hall, K. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of EMNLP 2011*, 62–72.

Mikolov, T.; Le, Q. V.; and Sutskever, I. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* .

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NeurIPS*, 3111–3119.

Mrkšić, N.; Vulić, I.; Séaghdha, D. Ó.; Leviant, I.; Reichart, R.; Gašić, M.; Korhonen, A.; and Young, S. J. 2017. Semantic Specialization of Distributional Word Vector Spaces

using Monolingual and Cross-Lingual Constraints. *Transactions of the Association for Computational Linguistics* 5: 309–324.

Nair, V.; and Hinton, G. E. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of ICML 2010*, 807–814.

Nakashole, N.; and Flauger, R. 2018. Characterizing departures from linearity in word translation. In *Proceedings of ACL 2018*. URL https://www.aclweb.org/anthology/P18-2036.

Newman-Griffis, D.; Lai, A. M.; and Fosler-Lussier, E. 2017. Insights into Analogy Completion from the Biomedical Domain. In *Proceedings of BioNLP 2017*, 19–28. URL https://www.aclweb.org/anthology/W17-2303.

Patra, B.; Moniz, J. R. A.; Garg, S.; Gormley, M. R.; and Neubig, G. 2019. Bilingual Lexicon Induction with Semi-supervision in Non-Isometric Embedding Spaces. In *Proceedings of ACL 2019*, 184–193. doi:10.18653/v1/p19-1018.

Peng, X.; Lin, C.; Stevenson, M.; and li, C. 2020. Revisiting the linearity in cross-lingual embedding mappings: From a perspective of word analogies. *arXiv preprint arXiv:2004.01079* .

Pires, T.; Schlinger, E.; and Garrette, D. 2019. How multilingual is Multilingual BERT? In *Proceedings of ACL 2019*, 4996–5001. URL http://arxiv.org/abs/1906.01502.

Ponti, E. M.; O'Horan, H.; Berzak, Y.; Vulić, I.; Reichart, R.; Poibeau, T.; Shutova, E.; and Korhonen, A. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics* 45(3): 559–601. URL https://arxiv.org/pdf/1807.00914.pdf.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP-IJCNLP 2019*, 3973–3983.

Rogers, A.; Drozd, A.; and Li, B. 2017. The (too Many) Problems of Analogical Reasoning with Word Vectors. In *Proceedings of StarSEM 2017*, 135–148.

Ruder, S.; Vulić, I.; and Søgaard, A. 2019. A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research* 65: 569–631. URL http://arxiv.org/abs/1706.04902.

Schluter, N. 2018. The word analogy testing caveat. In *Proceedings of NAACL 2018*, 242–246.

Schuster, T.; Ram, O.; Barzilay, R.; and Globerson, A. 2019. Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing. In *Proceedings of NAACL 2019*, 1599–1613.

Singh, J.; McCann, B.; Socher, R.; and Xiong, C. 2019. BERT is Not an Interlingua and the Bias of Tokenization. In *Proceedings of DeepLo@EMNLP-IJCNLP 2019*, 47–55.

Snyder, B.; Naseem, T.; and Barzilay, R. 2009. Unsupervised multilingual grammar induction. In *Proceedings of ACL 2009*, 73–81.

Søgaard, A.; Ruder, S.; and Vulić, I. 2018. On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of ACL 2018*, 778–788.

Täckström, O.; Das, D.; Petrov, S.; McDonald, R.; and Nivre, J. 2013. Token and Type Constraints for Cross-Lingual Part-of-Speech Tagging. *Transactions of the Association for Computational Linguistics* 1: 1–12.

Vulić, I.; Glavaš, G.; Reichart, R.; and Korhonen, A. 2019. Do We Really Need Fully Unsupervised Cross-Lingual Embeddings ? In *Proceedings of EMNLP-IJCNLP 2019*, 4407–4418.

Vulić, I.; Glavaš, G.; Mrkšić, N.; and Korhonen, A. 2018. Post-Specialisation: Retrofitting Vectors of Words Unseen in Lexical Resources. In *Proceedings of NAACL-HLT 2018*, 516–527.

Wu, S.; Conneau, A.; Li, H.; Zettlemoyer, L.; and Stoyanov, V. 2020. Emerging Cross-lingual Structure in Pretrained Language Models. In *Proceedings of ACL 2020*. URL http://arxiv.org/abs/1911.01464.

Wu, S.; and Dredze, M. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of EMNLP-IJCNLP 2019*, 833–844. URL http://arxiv.org/abs/1904.09077.

Yuan, M.; Zhang, M.; Durme, B. V.; Findlater, L.; and Boyd-Graber, J. L. 2020. Interactive Refinement of Cross-Lingual Word Embeddings. In *EMNLP*.

Zhang, M.; Fujinuma, Y.; Paul, M. J.; and Boyd-Graber, J. L. 2020. Why Overfitting Isn't Always Bad: Retrofitting Cross-Lingual Word Embeddings to Dictionaries. *ArXiv* abs/2005.00524.

Zhang, M.; Xu, K.; Kawarabayashi, K.-i.; Jegelka, S.; and Boyd-Graber, J. 2019. Are Girls Neko or Shōjo? Cross-Lingual Alignment of Non-Isomorphic Embeddings with Iterative Normalization. In *Proceedings of ACL 2019*, 3180–3189. URL http://arxiv.org/abs/1906.01622.

Zweigenbaum, P.; Sharoff, S.; and Rapp, R. 2017. Overview of the Third BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, 60–67.

Acknowledgements

We thank Angeliki Lazaridou, Phil Blunsom, Katja Filippova, and the reviewers for their thoughtful comments and suggestions. In order to retrieve the analogy data, we used SLING, developed by Michael Ringgaard. We thank Michael for his technical (and moral) support. Nicolas is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). Mareike Hartmann is funded by the Lundbeck Foundation. The work of Ivan Vulić is supported by the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (no 648909). Anders Søgaard is funded by a Google Focused Research Award.