Understanding Linearity of Cross-Lingual Word Embedding Mappings

Anonymous authors
Paper under double-blind review

Abstract

The technique of Cross-Lingual Word Embedding (CLWE) plays a fundamental role in tackling Natural Language Processing challenges for low-resource languages. Its dominant approaches assumed that the relationship between embeddings could be represented by a linear mapping, but there has been no exploration of the conditions under which this assumption holds. Such a research gap becomes very critical recently, as it has been evidenced that relaxing mappings to be non-linear can lead to better performance in some cases. We, for the first time, present a theoretical analysis that identifies the preservation of analogies encoded in monolingual word embeddings as a *necessary and sufficient* condition for the ground-truth CLWE mapping between those embeddings to be linear. On a novel cross-lingual analogy dataset[1] that covers five representative analogy categories for twelve distinct languages, we carry out experiments which provide direct empirical support for our theoretical claim. These results offer additional insight into the observations of other researchers and contribute inspiration for the development of more effective cross-lingual representation learning strategies.

1 Introduction

Cross-Lingual Word Embedding (CLWE) methods encode words from two or more languages in a shared high-dimensional space in which vectors representing lexical items with similar meanings (regardless of language) are closely located. Compared with alternative techniques, such as cross-lingual pre-trained language models, CLWE is orders of magnitude more efficient in terms of training corpora[2] and computational power requirements[3]. As a result, the topic has received significant attention as a promising means to support Natural Language Processing (NLP) for low-resource languages (including ancient languages) and has been used for a range of applications, e.g., Machine Translation (Herold et al., 2021), Sentiment Analysis (Sun et al., 2021), Question Answering (Zhou et al., 2021) and Text Summarisation (Peng et al., 2021).

The most successful CLWE approach, CLWE alignment, learns mappings between independently trained monolingual word vectors with very little, or even no, cross-lingual supervision (Ruder et al., 2019). One of the key challenges of these algorithms is the design of mapping functions. Motivated by the observation that word embeddings for different languages tend to be similar in structure (Mikolov et al., 2013b), many researchers have assumed that the mappings between cross-lingual word vectors are linear (Faruqui & Dyer, 2014; Lample et al., 2018b; Li et al., 2021).

Although models based on this assumption have demonstrated strong performance, it has recently been questioned. Researchers have claimed that the structure of multilingual word embeddings may not always be similar (Søgaard et al., 2018; Dubossarsky et al., 2020; Vulić et al., 2020), which led to the emergence

[1] This dataset and our code will be made publicly available upon the acceptance of this manuscript.

[2] For example, Kim et al. (2020) show that inadequate monolingual data size (fewer than one million *sentences*) is likely to lead to collapsed performance of XLM (Lample & Conneau, 2019) even for etymologically close language pairs. Meanwhile, CLWE can easily align word embeddings for languages such as African Amharic and Tigrinya for which only have millions of *tokens* (Zhang et al., 2020) are available.

[3] For example, XLM-R (Conneau et al., 2020) was trained on 500× Tesla V100 GPUs, whereas the training of VecMap (Artetxe et al., 2018) can be finished within minutes on a single Titan Xp GPU.

of approaches relaxing the mapping linearity (Glavaš & Vulić, 2020; Wang et al., 2021a) or using non-linear functions (Mohiuddin et al., 2020; Ganesan et al., 2021). These new methods can sometimes outperform the traditional linear counterparts, causing a debate around the suitability, or otherwise, of linear mappings. However, to the best of our knowledge, the majority of previous CLWE work has focused on empirical findings, and there has been no in-depth analysis of the conditions for the linearity assumption.

This paper approaches the problem from a novel perspective by establishing a link between the linearity of CLWE mappings and the preservation of encoded monolingual analogies. Our work is motivated by the observation that word analogies can be solved via the composition of semantics based on vector arithmetic (Mikolov et al., 2013c) and such linguistic regularities might be transferable across languages. More specifically, we notice that if analogies encoded in the embeddings of one language also appear in the embeddings of another, the corresponding multilingual vectors tend to form similar shapes (see Fig. 1), suggesting the CLWE mapping between them should be approximately linear. In other words, we suspect that the preservation of analogy encoding indicates the linearity of CLWE mappings.

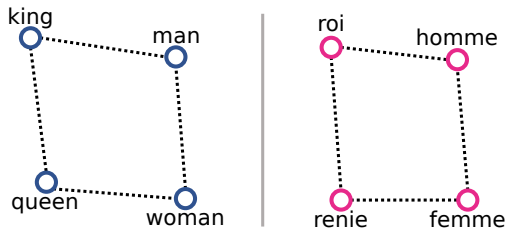

Figure 1: Wiki vectors (see § 4.3) of English (left) and French (right) analogy word pairs based on PCA (Wold et al., 1987). NB: We manually rotate the visualisation to highlight structural similarity.

Our hypothesis is verified both theoretically and empirically. We make a justification that the preservation of analogy encoding should be a *sufficient and necessary* condition for the linearity of CLWE mappings. To provide empirical validation, we first define indicators to qualify the linearity of the ground-truth CLWE mapping ($\mathcal{S}_{\text{LMP}}$) and its preservation of analogy encoding ($\mathcal{S}_{\text{PAE}}$). Next, we build a novel cross-lingual word analogy corpus containing five analogy categories (both semantic and syntactic) for twelve languages that pose pairs of diverse etymological distances. We then benchmark $\mathcal{S}_{\text{LMP}}$ and $\mathcal{S}_{\text{PAE}}$ on three representative series of word embeddings. In all setups tested, we observe a significant correlation between $\mathcal{S}_{\text{LMP}}$ and $\mathcal{S}_{\text{PAE}}$, which provides empirical support for our hypothesis. With this insight, we offer explanations to why the linearity assumption occasionally fails, and consequently, discuss how our research can benefit the development of more effective CLWE algorithms. We also recommend the use of $\mathcal{S}_{\text{PAE}}$ to assess mapping linearity in CLWE applications.

This paper's contributions are summarised as: (1) Introduces the previously unnoticed relationship between the linearity of CLWE mappings and the preservation of encoded word analogies. (2) Provides a theoretical analysis of this relationship. (3) Describes the construction of a novel cross-lingual analogy test set with five categories of word pairs aligned across twelve diverse languages. (4) Provides empirical evidence of our claim and introduces $\mathcal{S}_{\text{PAE}}$ to estimate the analogy encoding preservation (and therefore the mapping linearity). (5) Discusses implications of these results, regarding the interpretation of previous results and as well as the future development of cross-lingual representations.

2 Related Work

Linearity of CLWE Mapping. Mikolov et al. (2013b) discovered that the vectors of word translations exhibit similar structures across different languages. Researchers made use of this by assuming that mappings between multilingual embeddings could be modelled using simple linear transformations. This framework turned out to be effective in numerous studies which demonstrated that linear mappings are able to produce accurate CLWEs with weak or even no supervision (Artetxe et al., 2017; Lample et al., 2018b; Artetxe et al., 2018; Wang et al., 2020; Li et al., 2021).

One way in which this is achieved is through the application of a normalisation technique called "mean centring", which (for each language) subtracts the average monolingual word vector from all word embeddings, so that this mean vector becomes the origin of the vector space (Xing et al., 2015; Artetxe et al., 2016;

Ruder et al., 2019). This step has the effect of simplifying the mapping from being *affine* (i.e., equivalent to a shifting operation plus a linear mapping) to *linear* by removing the shifting operation.

However, recent work has cast doubt on this linearity assumption, leading researchers to experiment with the use of non-linear mappings. Nakashole & Flauger (2018) and Wang et al. (2021a) pointed out that structural similarities may only hold across particular regions of the embedding spaces rather than over their entirety. Søgaard et al. (2018) examined word vectors trained using different corpora, models and hyper-parameters, and concluded configuration dissimilarity between the monolingual embeddings breaks the assumption that the mapping between them is linear. Patra et al. (2019) investigated various language pairs and discovered that a higher etymological distance is associated with degraded the linearity of CLWE mappings. Vulić et al. (2020) additionally argued that factors such as limited monolingual resources may also weaken the linearity assumption.

These findings motivated work on designing non-linear mapping functions in an effort to improve CLWE performance. For example, Nakashole (2018) and Wang et al. (2021a) relaxed the linearity assumption by combining multiple linear CLWE mappings; Patra et al. (2019) developed a semi-supervised model that loosened the linearity restriction; Lubin et al. (2019) attempted to reduce the dissimilarity between multilingual embedding manifolds by refining learnt dictionaries; Glavaš & Vulić (2020) first trained a globally optimal linear mapping, then adjusted vector positions to achieve better accuracy; Mohiuddin et al. (2020) used two independently pre-trained auto-encoders to introduce non-linearity to CLWE mappings; Ganesan et al. (2021) obtained inspirations via the back translation paradigm, hence framing CLWE training as to explicitly solve a non-linear and bijective transformation between multilingual word embeddings. Despite these non-linear mappings outperforming their linear counterparts in many setups, in some settings the linear mappings still seem more successful, e.g., the alignment between Portuguese and English word embeddings in Ganesan et al. (2021). Moreover, training non-linear mappings is typically more complex and thus requires more computational resources.

Albeit at the significant recent attention to this problem by the research community, it is still unclear under what condition the linearity of CLWE mappings holds. This paper makes the first attempt to close this research gap by providing both theoretical and empirical contributions.

Analogy Encoding. Analogy is a fundamental concept within cognitive science (Gentner, 1983) that has received significant focus from the NLP community, since the observation that it can be represented using word embeddings and vector arithmetic (Mikolov et al., 2013c). A popular example based on the analogy "*king is to man as queen is to woman*" shows that the vectors representing the four terms ($\boldsymbol{x_{king}}$, $\boldsymbol{x_{man}}$, $\boldsymbol{x_{queen}}$ and $\boldsymbol{x_{woman}}$) exhibit the following relation:

$$\boldsymbol{x_{king}} - \boldsymbol{x_{man}} \approx \boldsymbol{x_{queen}} - \boldsymbol{x_{woman}}. \tag{1}$$

Since this discovery, the task of analogy completion has commonly been employed to evaluate the quality of pre-trained word embeddings (Mikolov et al., 2013c; Pennington et al., 2014; Levy & Goldberg, 2014a). This line of research has directly benefited downstream applications (e.g., representation bias removal (Prade & Richard, 2021)) and other relevant domains (e.g., automatic knowledge graph construction (Wang et al., 2021b)). Theoretical analysis has demonstrated a link between embeddings' analogy encoding and the Pointwise Mutual Information of the training corpus (Arora et al., 2016; Gittens et al., 2017; Allen & Hospedales, 2019; Ethayarajh et al., 2019; Fournier & Dunbar, 2021). Nonetheless, as far as we are aware, the connection between the preservation of analogy encoding and the linearity of CLWE mappings has not been previously investigated.

3 Theoretical Basis

We denote a ground-truth CLWE mapping as $\mathcal{M} : \mathbf{X} \rightarrow \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are monolingual word embeddings independently trained for languages $\mathrm{L_X}$ and $\mathrm{L_Y}$, respectively.

Proposition. Encoded analogies are preserved during the CLWE mapping $\mathcal{M} \Longleftrightarrow \mathcal{M}$ is affine.

Remarks. Following Eq. (1), the preservation of analogy encoding under a mapping can be formalised as

$$\boldsymbol{x_\alpha} - \boldsymbol{x_\beta} = \boldsymbol{x_\gamma} - \boldsymbol{x_\theta} \implies \mathcal{M}(\boldsymbol{x_\alpha}) - \mathcal{M}(\boldsymbol{x_\beta}) = \mathcal{M}(\boldsymbol{x_\gamma}) - \mathcal{M}(\boldsymbol{x_\theta}), \tag{2}$$

where $\boldsymbol{x_\alpha}, \boldsymbol{x_\beta}, \boldsymbol{x_\gamma}, \boldsymbol{x_\theta} \in \mathbf{X}$.

If $\mathcal{M}$ is affine, for d-dimensional monolingual embeddings X we have

$$\mathcal{M}(\boldsymbol{x}) \coloneqq M\boldsymbol{x} + \boldsymbol{b}, \tag{3}$$

where $x \in X$, $M \in \mathbb{R}^{d\times d}$, and $\boldsymbol{b} \in \mathbb{R}^{d\times 1}$ ($\mathbb{R}$ is the set of real numbers).

Proof: **Eq.** (2) $\implies$ **Eq.** (3) **(i.e., the forward implication).** To begin with, by adopting the mean centring operation in § 2, we shift the coordinates of the space of $\mathbf{X}$, ensuring

$$\mathcal{M}(\vec{0}) = \vec{0}. \tag{4}$$

This step greatly simplifies the derivations afterwards, because from now on we just need to demonstrate that $\mathcal{M}$ is a *linear mapping*, i.e., it can be written as $M\boldsymbol{x}$. By definition, this is equivalent to showing that $\mathcal{M}$ preserves both the operations of addition (a.k.a. additivity) and scalar multiplication (a.k.a. homogeneity).

Additivity can be proved by observing that $(\boldsymbol{x_i} + \boldsymbol{x_j}) - \boldsymbol{x_j} = \boldsymbol{x_i} - \vec{0}$ and therefore,

$$\begin{aligned}(\boldsymbol{x_i} + \boldsymbol{x_j}) - \boldsymbol{x_j} = \boldsymbol{x_i} - \vec{0} &\xRightarrow{\text{Eq. (2)}} \mathcal{M}(\boldsymbol{x_i} + \boldsymbol{x_j}) - \mathcal{M}(\boldsymbol{x_j}) = \mathcal{M}(\boldsymbol{x_i}) - \mathcal{M}(\vec{0}) \\ &\xRightarrow{\text{Eq. (4)}} \mathcal{M}(\boldsymbol{x_i} + \boldsymbol{x_j}) = \mathcal{M}(\boldsymbol{x_i}) + \mathcal{M}(\boldsymbol{x_j}).\end{aligned} \tag{5}$$

Homogeneity needs a proof that seems more complex, which consists of four steps.

• **Step 1**: Observe that $\vec{0} - \boldsymbol{x_i} = -\boldsymbol{x_i} - \vec{0}$, similar to Eq. (5) we can show that

$$\begin{aligned}\vec{0} - \boldsymbol{x_i} = -\boldsymbol{x_i} - \vec{0} &\xRightarrow{\text{Eq. (2)}} \mathcal{M}(\vec{0}) - \mathcal{M}(\boldsymbol{x_i}) = \mathcal{M}(-\boldsymbol{x_i}) - \mathcal{M}(\vec{0}) \\ &\xRightarrow[\times(-1)]{\text{Eq. (4)}} \mathcal{M}(\boldsymbol{x_i}) = -\mathcal{M}(-\boldsymbol{x_i}).\end{aligned} \tag{6}$$

• **Step 2**: Using *mathematical induction*, for arbitrary $\boldsymbol{x_i}$, we show that

$$\forall m \in \mathbb{N}^+,\ \mathcal{M}(m\boldsymbol{x_i}) = m\mathcal{M}(\boldsymbol{x_i}) \tag{7}$$

holds, where $\mathbb{N}^+$ is the set of positive natural numbers, as
Base Case: Trivially holds when $m = 1$.
Inductive Step: Assume the inductive hypothesis that $m = k$ ($k \in \mathbb{N}^+$), i.e.,

$$\mathcal{M}(k\boldsymbol{x_i}) = k\mathcal{M}(\boldsymbol{x_i}). \tag{8}$$

Then, as required, when $m = k + 1$,

$$\mathcal{M}\big((k+1)\boldsymbol{x_i}\big) \overset{\text{Eq. (5)}}{=\!=} \mathcal{M}(k\boldsymbol{x_i}) + \mathcal{M}(\boldsymbol{x_i}) \overset{\text{Eq. (8)}}{=\!=} k\mathcal{M}(\boldsymbol{x_i}) + \mathcal{M}(\boldsymbol{x_i}) = (k+1)\mathcal{M}(\boldsymbol{x_i}).$$

• **Step 3**: We further justify that

$$\forall n \in \mathbb{N}^+,\ \mathcal{M}(\frac{\boldsymbol{x_i}}{n}) = \frac{\mathcal{M}(\boldsymbol{x_i})}{n}, \tag{9}$$

which, due to Eq. (4), trivially holds when $n = 1$; as for $n > 1$,

$$\begin{aligned}\mathcal{M}(\frac{\boldsymbol{x_i}}{n}) = \mathcal{M}\big(\boldsymbol{x_i} + (-\frac{n-1}{n}\boldsymbol{x_i})\big) &\overset{\text{Eq. (5)}}{=\!=} \mathcal{M}(\boldsymbol{x_i}) + \mathcal{M}(-\frac{n-1}{n}\boldsymbol{x_i}) \\ &\overset{\text{Eq. (6)}}{=\!=} \mathcal{M}(\boldsymbol{x_i}) - \mathcal{M}(\frac{n-1}{n}\boldsymbol{x_i}) \overset{\text{Eq. (7)}}{=\!=} \mathcal{M}(\boldsymbol{x_i}) - (n-1)\mathcal{M}(\frac{\boldsymbol{x_i}}{n})\end{aligned}$$

directly yields $\mathcal{M}(\frac{\boldsymbol{x_i}}{n}) = \frac{\mathcal{M}(\boldsymbol{x_i})}{n}$, i.e., Eq. (9).

• **Step 4**: Considering the set of rational numbers $\mathbb{Q} = \{0\} \cup \{\pm\frac{m}{n}|\forall m, n\}$, Eqs. (4), (6), (7) and (9) jointly justifies the homogeneity of $\mathcal{M}$ for $\mathbb{Q}$. Because $\mathbb{Q} \subset \mathbb{R}$ is a *dense set*, homogeneity of $\mathcal{M}$ also holds over $\mathbb{R}$, see Kleiber & Pervin (1969).

Finally, combined with the additivity that has been already justified above, linearity of CLWE mapping $\mathcal{M}$ is proved, i.e., Eq. (2) $\Longrightarrow$ Eq. (3). □

Proof: **Eq.** (3) $\Longrightarrow$ **Eq.** (2) **(i.e., the reverse direction).** Justifying this direction is quite straightforward:

$$\begin{aligned}\boldsymbol{x_\alpha} - \boldsymbol{x_\beta} = \boldsymbol{x_\gamma} - \boldsymbol{x_\theta} &\Longrightarrow M\boldsymbol{x_\alpha} - M\boldsymbol{x_\beta} = M\boldsymbol{x_\gamma} - M\boldsymbol{x_\theta} \\ &\Longrightarrow M\boldsymbol{x_\alpha} + \boldsymbol{b} - (M\boldsymbol{x_\beta} + \boldsymbol{b}) = M\boldsymbol{x_\gamma} + \boldsymbol{b} - (M\boldsymbol{x_\theta} + \boldsymbol{b}) \\ &\Longrightarrow \mathcal{M}(\boldsymbol{x_\alpha}) - \mathcal{M}(\boldsymbol{x_\beta}) = \mathcal{M}(\boldsymbol{x_\gamma}) - \mathcal{M}(\boldsymbol{x_\theta}).\end{aligned}$$ □

Summarising the proofs for both the forward and reverse directions, we conclude that the proposition holds.

Please note, the high-level assumption of our derivations is that word embedding spaces can be treated as continuous vector spaces, an assumption commonly adopted in previous work, e.g., Levy & Goldberg (2014b), Hashimoto et al. (2016), Zhang et al. (2018), and Ravfogel et al. (2020). Nevertheless, we argue that the inherent discreteness of word embeddings should not be ignored. The following sections complement this theoretical insight via experiments which confirm the claim holds empirically.

4 Experiment

Our experimental protocol assesses the linearity of the mapping between each pair of pre-trained monolingual word embeddings. We also quantify the extent to which this mapping preserves encoded analogies, i.e., satisfies the condition of Eq. (2). We then analyse the correlation between these two indicators. A strong correlation provides evidence to support our theory, and *vice versa*. The indicators used are described in § 4.1. Unfortunately, there are no suitable publicly available corpora for our proposed experiments, so we develop a novel word-level analogy test set that is fully parallel across languages, namely xANLG (see § 4.2). The pre-trained embeddings used for the tests are described in § 4.3.

4.1 Indicators

4.1.1 Linearity of CLWE Mapping

Direct measurement of the linearity of a ground-truth CLWE mapping is challenging. One relevant approach is to benchmark the similarity (sometimes described as "isomorphism" in previous work) between multilingual word embedding, where the mainstream and state-of-the-art indicators are the so-called spectral-based algorithms (Søgaard et al., 2018; Dubossarsky et al., 2020). However, such methods assume the number of tested vectors to be much larger than the number of dimensions, which does not apply in our scenario (see § 4.2). Therefore, we choose to evaluate linearity via the goodness-of-fit of the optimal linear CLWE mapping, which is measured as

$$\mathcal{S}_{\text{LMP}} := -||M^\star X - Y||_F \quad \text{with} \quad M^\star = \arg\min_M ||MX - Y||_F,$$

where $||\cdot||_F$ denotes the Frobenius norm. To obtain matrices X and Y, from $\mathbf{X}$ and $\mathbf{Y}$ respectively, we first retrieve the vectors corresponding to lexicons of a ground-truth L_X-L_Y dictionary and concatenate them into two matrices. More specifically, if two vectors (represented as rows) share the same index in the two matrices (one for each language), their corresponding words form a translation pair, i.e., the rows of these matrices are aligned. Then "mean centring" is applied to satisfy Eq. (4) and the matrices are normalised by

scaling their Frobenius norm to 1 for fair comparisons across different mapping pairs. Gradient descent is applied to find $M^\star$, the matrix that determines the optimal linear mapping (Mikolov et al., 2013b).

Large absolute values of $\mathcal{S}_{\text{LMP}}$ mean that the optimal linear mapping is an accurate model of the true relationship between the embeddings, and *vice versa*. $\mathcal{S}_{\text{LMP}}$ therefore indicates the degree to which CLWE mappings are linear.

4.1.2 Preservation of Analogy Encoding

To assess how well analogies are preserved across embeddings, we start by probing how analogies are encoded in the monolingual word embeddings. We use the set-based LRCos, the state-of-the-art analogy mining tool for static word embeddings (Drozd et al., 2016).[4] It provides a score in the range of 0 to 1, indicating the correctness of analogy completion in a single language. For the extension in a cross-lingual setup, we further compute the geometric mean:

$$\mathcal{S}_{\text{PAE}} \coloneqq \sqrt{\text{LRCos}(\mathbf{X}) \times \text{LRCos}(\mathbf{Y})},$$

where LRCos$(\cdot)$ is the accuracy of analogy completion provided by LRCos for embedding $\mathbf{X}$. To simplify our discussion and analysis from now onward, when performing CLWE mappings, by default we select the monolingual embeddings that best encode analogy, i.e., we restrict LRCos$(\mathbf{X}) \geq$ LRCos$(\mathbf{Y})$. $\mathcal{S}_{\text{PAE}} = 1$ indicates all analogies are well encoded in both embeddings, and are preserved by the ground-truth mapping between them. On the other hand, lower $\mathcal{S}_{\text{PAE}}$ values indicate deviation from the condition of Eq. (2).

4.1.3 Validity of $\mathcal{S}_{\text{PAE}}$

As an aide, we explore the properties of the $\mathcal{S}_{\text{PAE}}$ indicator to demonstrate its robustness for the interested reader. The score produced by LRCos is relative to a pre-specified set of *known* analogies. In theory, a low LRCos$(\mathbf{X})$ score may not reliably indicate that $\mathbf{X}$ does not encode analogies well since there may be other word pairings within that set that produce higher scores. This naturally raises a question: *does $\mathcal{S}_{\text{PAE}}$ really promise the validity as the indicator of analogy encoding preservation?* In other words, it is necessary to investigate whether there exists an *unknown* analogy word set encoded by the tested embeddings to an equal or higher degree. If there is, then $\mathcal{S}_{\text{PAE}}$ may not reflect the preservation of analogy encoding completely, as unmatched analogy test sets may lead to low LRCos scores even for monolingual embeddings that encode analogies well. We demonstrate that the problem can be considered as an optimal transportation task and $\mathcal{S}_{\text{PAE}}$ is guaranteed to be a reliable indicator.

As analysed by Ethayarajh et al. (2019), the degree to which word pairs are encoded as analogies in word embeddings is equivalent to the likelihood that the end points of any two corresponding vector pairs form a high-dimensional coplanar parallelogram. More formally, this task is to identify

$$\mathbf{P}^\star = \arg\min_{\mathbf{P}} \sum_{\boldsymbol{x} \in \mathbf{X}} \mathcal{C}\big(\mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x})\big), \tag{10}$$

where $\mathbf{P}$ is one possible pairing of vectors in $\mathbf{X}$ and $\mathcal{C}(\cdot)$ is the cost of a given transportation scheme. $\mathcal{T}^{\mathbf{P}}_{\square}(\cdot)$ denotes the corresponding cost-optimal process of moving vectors to satisfy

$$\begin{aligned} &\forall\{(\boldsymbol{x_\alpha}, \boldsymbol{x_\beta}), (\boldsymbol{x_\gamma}, \boldsymbol{x_\theta})\} \subseteq \mathbf{P}, \\ &\quad \mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x_\alpha}) - \mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x_\beta}) = \mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x_\gamma}) - \mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x_\theta}), \end{aligned} \tag{11}$$

i.e., the end points of $\mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x_\alpha})$, $\mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x_\beta})$, $\mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x_\gamma})$ and $\mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x_\theta})$ form a parallelogram.

Therefore, in each language and analogy category of xANLG, we first repeatedly redo the vector pairing to traverse *all* $\mathbf{P}$. Next, we need to obtain $\mathcal{T}^{\mathbf{P}}_{\square}(\cdot)$ that minimises $\sum_{\boldsymbol{x} \in \mathbf{X}} \mathcal{C}\big(\mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x})\big)$ in Eq. (10). Our algorithm is explained using the example in Fig. 2, where the cardinality of $\mathbf{X}$ and $\mathbf{P}$ is 8 and 4, respectively.

[4] We have tried alternatives including 3CosAdd (Mikolov et al., 2013a), PairDistance (Levy & Goldberg, 2014a) and 3CosMul (Levy et al., 2015), verifying that they are less accurate than LRCos in most cases. Still, in the experiments they all exhibit similar trends as shown in Tab. 2.

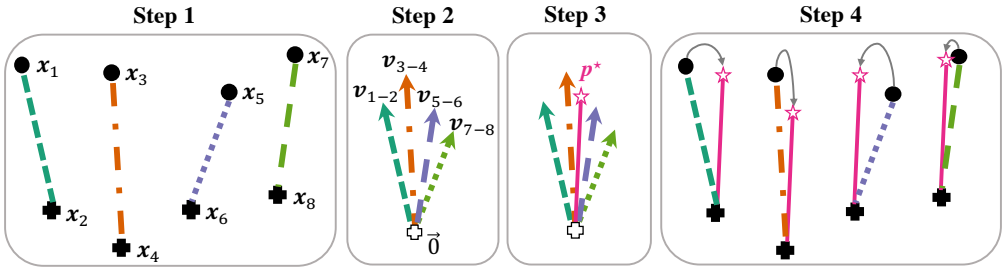

Figure 2: An example of solving $\mathcal{T}^{\mathbf{P}}_{\square}(\cdot)$ in Eq. (11), with $\mathbf{P} = \{(\boldsymbol{x}_1, \boldsymbol{x}_2), (\boldsymbol{x}_3, \boldsymbol{x}_4), (\boldsymbol{x}_5, \boldsymbol{x}_6), (\boldsymbol{x}_7, \boldsymbol{x}_8)\}$. In the figure we adjust the position of $\boldsymbol{x}_1$, $\boldsymbol{x}_3$, $\boldsymbol{x}_5$ and $\boldsymbol{x}_7$ in the last step, but it is worth noting that there also exists other feasible $\mathcal{T}^{\mathbf{P}}_{\square}(\cdot)$ given $\boldsymbol{p}^\star$, e.g., to tune $\boldsymbol{x}_2$, $\boldsymbol{x}_4$, $\boldsymbol{x}_6$ and $\boldsymbol{x}_8$ instead.

- **Step 1**: Link the end points of the vectors within each word pair, hence our target is to adjust these end points so that all connecting lines not only have equal length but also remain parallel.
- **Step 2**: For each vector pair $(\boldsymbol{x}_{\boldsymbol{\alpha}}, \boldsymbol{x}_{\boldsymbol{\beta}}) \in \mathbf{P}$, vectorise its connecting line into an offset vector as $\boldsymbol{v}_{\boldsymbol{\alpha}-\boldsymbol{\beta}} = \boldsymbol{x}_{\boldsymbol{\alpha}} - \boldsymbol{x}_{\boldsymbol{\beta}}$.
- **Step 3**: As the start points of all such offset vectors are aggregated at $\vec{0}$, seek a vector $\boldsymbol{p}^\star$ that minimises the total transportation cost between the end point of $\boldsymbol{p}^\star$ and those of all offset vectors (again, note they share a start point at $\vec{0}$).
- **Step 4**: Perform the transportation so that all offset vectors become $\boldsymbol{p}^\star$, i.e.,

$$\forall (\boldsymbol{x}_{\boldsymbol{\alpha}}, \boldsymbol{x}_{\boldsymbol{\beta}}) \in \mathbf{P},\ \mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x}_{\boldsymbol{\alpha}}) - \mathcal{T}^{\mathbf{P}}_{\square}(\boldsymbol{x}_{\boldsymbol{\beta}}) = \boldsymbol{p}^\star.$$

In this way, the tuned vector pairs can always form perfect parallelograms. Obviously, as $\boldsymbol{p}^\star$ is at the cost-optimal position (see Step 3), this vector-adjustment scheme is also cost-optimal.

Solving $\boldsymbol{p}^\star$ for high dimensions is non-trivial in real world and is a special case of the NP-hard Facility Location Problem (a.k.a. the P-Median Problem) (Kariv & Hakimi, 1979). We, therefore, use the `scipy.optimize.fmin` implementation of the Nelder-Mead simplex algorithm (Nelder & Mead, 1965) to provide a good-enough solution. We experimented with implementing $\mathcal{C}(\cdot)$ using Euclidean, Taxicab and Cosine distances. For all analogy categories in all languages, $\mathbf{P}^\star$ coincides perfectly with the pre-defined pairing of xANLG. This analysis provides evidence that the situation where "an unknown kind of analogy is better encoded than the ones used" does not occur in practice. $\mathcal{S}_{\text{PAE}}$ is thus trustworthy.

4.2 Datasets

Calculating the correlation between $\mathcal{S}_{\text{LMP}}$ and $\mathcal{S}_{\text{PAE}}$ requires a cross-lingual word analogy dataset. This resource would allow us to simultaneously (1) construct two aligned matrices X and Y to check the linearity of CLWE mappings, and (2) obtain the monolingual LRCos scores of both $\mathbf{X}$ and $\mathbf{Y}$. Three relevant resources were identified, although none of them is suitable for our study.

- Brychcín et al. (2019) described a cross-lingual analogy dataset consisting of word pairs from six closely related European languages, but it has never been made publicly available.
- Ulčar et al. (2020) open-sourced the MCIWAD dataset for nine languages, but the analogy words in different languages are not parallel[5].
- Garneau et al. (2021) produced the cross-lingual WiQueen dataset. Unfortunately, a large part of its entries are proper nouns or multi-word terms instead of single-item words, leading to low coverage on the vocabularies of embeddings.

Consequently, we develop xANLG, which we believe to be the first (publicly available) cross-lingual word analogy corpus. For consistency with previous work, xANLG is bootstrapped using established monolingual analogies and cross-lingual dictionaries. xANLG is constructed by starting with a *bilingual* analogy dataset, say, that for L_X and L_Y. Within each analogy category, we first translate word pairs of the L_X analogy

[5]Personal communication with the authors.

	Category	#	DE	EN	ES	FR	HI	PL
xANLG$_G$	CAP†	31	Budapest Ungarn	Budapest Hungary	Budapest Hungría	Budapest Hongrie	बुडापेस्ट हंगरी	Budapeszt Węgry
	GNDR†	30	sohn tochter	son daughter	hijo hija	fils fille	बेटा बेटी	syn córka
	NATL†	34	Peru Peruanisch	Peru Peruvian	Perú Peruano	Pérou Péruvien	पेरू पेरू	Peru Peruwiański
	G-PL‡	31	kind kinder	child children	niño niños	enfant enfants	बच्चा बच्चे	dziecko dzieci
	Category	**#**	**EN**	**ET**	**FI**	**HR**	**LV**	**RU**	**SL**
xANLG$_M$	ANIM†	32	eagle bird	kotkas lind	kotka lintu	orao ptica	ērglis putns	орёл птица	orel ptica
	G-PL‡	31	machine machines	masin masinad	kone koneet	stroj strojevi	mašīna mašīnas	машина машины	stroj stroji

Table 1: Summary of and examples from the xANLG corpus. # denotes the number of cross-lingual analogy word pairs in each language. †Semantic: `animal-species`|ANIM, `capital-world`|CAP, `male-female`|GNDR, `nation-nationality`|NATL. ‡Syntactic: `grammar-plural`|G-PL.

corpus into L_Y, using an available L_X-L_Y dictionary. Next, we check if any translation coincides with its original word pair in L_Y. If it does, such a word pair (in both L_X and L_Y) will be added into the bilingual dataset. This process is repeated for multiple languages to form a cross-lingual corpus.

We use the popular MUSE dictionary (Lample et al., 2018a) which contains a wide range of language pairs. Two existing collections of analogies are utilised. (1) **Google Analogy Test Set (GATS)** (Mikolov et al., 2013c), the *de facto* standard benchmark of embedding-based analogy solving. We adopt its extended English version, Bigger Analogy Test Set (BATS) (Gladkova et al., 2016), supplemented with several datasets in other languages inspired by the original GATS: French, Hindi and Polish (Grave et al., 2018), German (Köper et al., 2015) and Spanish (Cardellino, 2019). (2) The aforementioned **Multilingual Culture-Independent Word Analogy Datasets (MCIWAD)** (Ulčar et al., 2020).

Due to the differing characteristics of these datasets (e.g., the composition of analogy categories), they are used to produce two separate corpora: xANLG$_G$ and xANLG$_M$. Only categories containing at least 30 word pairs aligned across all languages in the dataset were included. For comparison, 60% of the semantic analogy categories in the commonly used GATS dataset contains fewer than 30 word pairs. The rationale for selecting this value was that it allows a reasonable number of analogy completion questions to be generated.[6] Information in xANLG$_G$ and xANLG$_M$ for the `capital-country` of Hindi was supplemented with manual translations by native speakers.

The xANLG dataset contains five distinct analogy categories, including both syntactic (morphological) and semantic analogies, and twelve languages from a diverse range of families (see Tab. 1). From Indo-European languages, one belongs to the Indo-Aryan branch (Hindi|HI), one to the Baltic branch (Latvian|LV), two to the Germanic branch (English|EN, German|DE), two to the Romance branch (French|FR, Spanish|ES) and four to the Slavonic branch (Croatian|HR, Polish|PL, Russian|RU, Slovene|SL). Two non-Indo-European languages, Estonian|ET and Finnish|FI, both from the Finnic branch of the Uralic family, are also included. In total, they form 15 and 21 languages pairs for xANLG$_G$ and xANLG$_M$, respectively. These pairs span multiple etymological combinations, i.e., intra-language-branch (e.g., ES-FR), inter-language-branch (e.g., DE-RU) and inter-language-family (e.g., HI-ET).

[6] 30 word pairs can be used to generate as many as 3480 unique analogy completion questions such as "*king*:*man* :: *queen*:?" (see Appendix A).

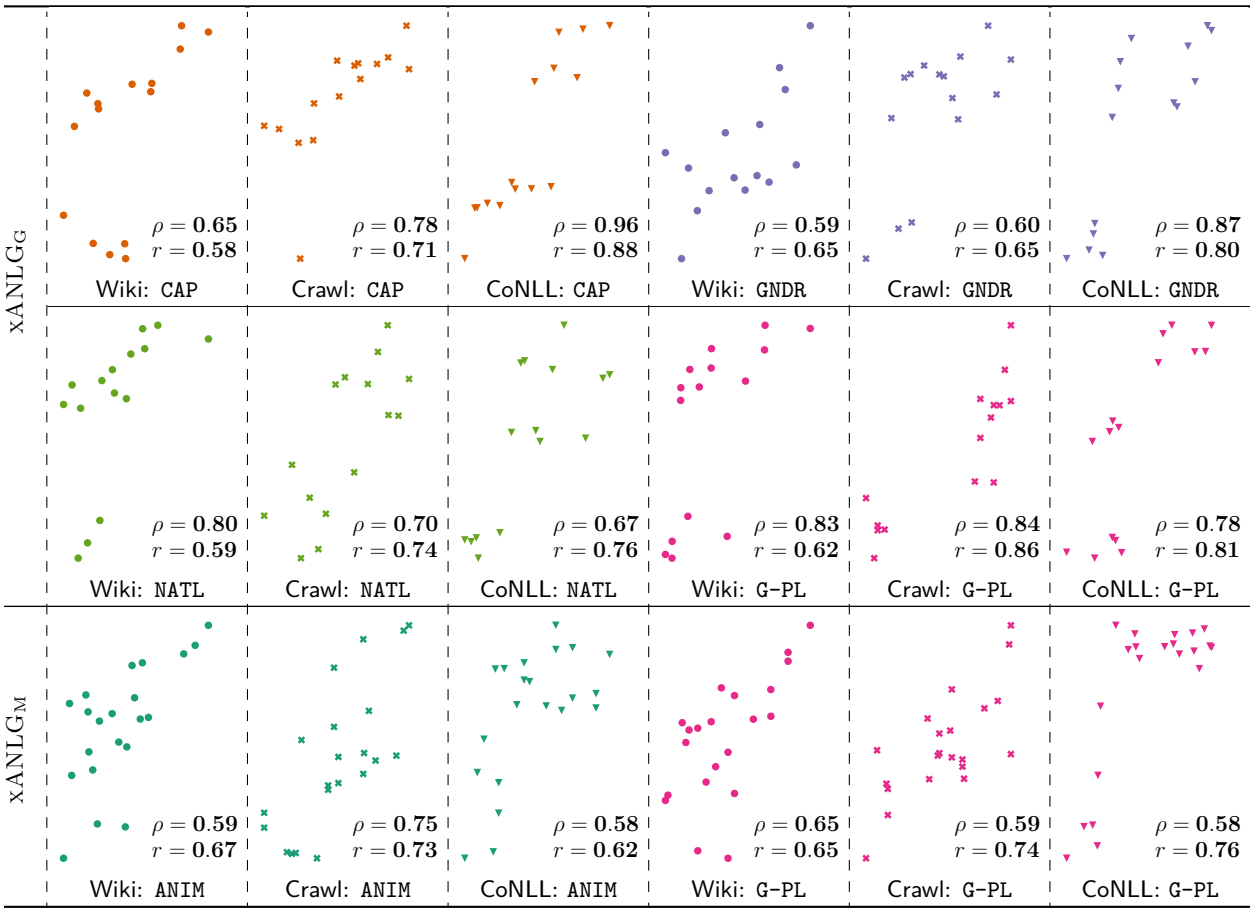

Table 2: Correlation coefficients (Spearman's ρ and Pearson's r) between $\mathcal{S}_{\text{LMP}}$ and $\mathcal{S}_{\text{PAE}}$. For all groups, we conduct significance tests to estimate the p-value. Empirically, the p-value is always less than 1e-2 (in most groups it is even less than 1e-3), indicating a very high confidence level for the experiment results. To facilitate future research and analyses, we present the raw $\mathcal{S}_{\text{LMP}}$ and LRCos data in Appendix B.

4.3 Word Embeddings

To cover the language pairs used in xANLG, we make use of static word embeddings pre-trained on the twelve languages used in the resource. These embeddings consist of three representative open-source series that employ different training corpora, are based on different embedding algorithms, and have different vector dimensions.

- **Wiki**[7]: 300-dimensional, trained on Wikipedia using the Skip-Gram version of FastText (refer to Bojanowski et al. (2017) for details).
- **Crawl**[8]: 300-dimensional, trained on CommonCrawl plus Wikipedia using FastText-CBOW.
- **CoNLL**[9]: 100-dimensional, trained on the CoNLL corpus (without lemmatisation) using Word2Vec (Mikolov et al., 2013c).

5 Result

Both Spearman's rank-order (ρ) and Pearson product-moment (r) correlation coefficients are computed to measure the correlation between $\mathcal{S}_{\text{LMP}}$ and $\mathcal{S}_{\text{PAE}}$. Note that, it is not possible to compute the correlations

[7] https://fasttext.cc/docs/en/pretrained-vectors.html
[8] https://fasttext.cc/docs/en/crawl-vectors.html
[9] http://vectors.nlpl.eu/repository/

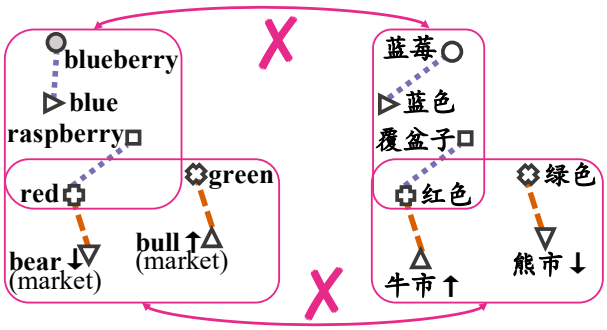

Figure 3: Illustration of example scenarios where the CLWE mapping is non-linear. Translations of English (left) and Chinese (right) terms are indicated by shared symbols. **Upper**: The vector for "*blueberry*" (shadowed) is ill-positioned in the embedding space, so the condition of Eq. (2) is no longer satisfied. **Lower**: In the financial domain some Eastern countries (e.g., China and Japan) traditionally use "*red*" to indicate growth and "*green*" for reduction, while Western countries (e.g., US and UK) assign the opposite meanings to these terms, also not satisfying the condition of Eq. (2).

between all pairs due to (1) the number of dimensions varies across embeddings series, and (2) the source and target embeddings have been pre-processed independently for different mappings. Instead, results are grouped by embedding method and analogy category.

Figures in Tab. 2 show that a significant positive correlation between $\mathcal{S}_{\text{PAE}}$ and $\mathcal{S}_{\text{LMP}}$ is observed for all setups. In terms of the Spearman's ρ, among the 18 groups, 5 exhibit *very strong* correlation ($\rho \geq 0.80$) (with a maximum at 0.96 for CoNLL embeddings on CAP of xANLG_{G}), 4 show *strong* correlation ($0.80 > \rho \geq 0.70$), and the others have *moderate* correlation ($0.70 > \rho \geq 0.50$) (with a minimum at 0.58: CoNLL embeddings on ANIM and G-PL of xANLG_{M}). Interestingly, although we do not assume a linear relationship in § 3, large values for the Pearson's r are obtained in practice. To be exact, 4 groups indicate very strong correlation, 6 have strong correlation, while others retain moderate correlation (the minimum r value is 0.58: Wiki embeddings on CAP and G-PL of xANLG_{G}). These results provide empirical evidence that supplements our theoretical analysis (§ 3) of the relationship between linearity of mappings and analogy preservation.

In addition, we explored whether the analogy type (i.e., semantic or syntactic) affects the correlation. To bootstrap the analysis, for both kinds of correlation coefficients, we divide the 18 experiment groups into two splits, i.e., 12 semantic ones and 6 syntactic ones. After that, we compute a two-treatment ANOVA (Fisher, 1925). For both Spearman's ρ and Pearson's r, the results are not significant at $p < 0.1$. Therefore, we conclude that the connection between CLWE mapping linearity and analogy encoding preservation holds across analogy types. We thus recommend testing $\mathcal{S}_{\text{PAE}}$ *before* implementing CLWE alignment as an indicator of whether a linear transformation is a good approximation of the ground-truth CLWE mapping.

6 Further Discussion

Prior work relevant to the linearity of CLWE mappings has largely been observational (see § 2). This section sheds new light on these past studies from the novel perspective of word analogies.

Explaining Non-Linearity. We provide two suggested reasons why CLWE mappings are sometimes not approximately linear, both linked with the condition of Eq. (2) not being met.

The first may be issues with individual monolingual embeddings (see one such example in the upper part of Fig. 3). In particular, popular word embedding algorithms lack the capacity to ensure semantic continuity over the entire embedding space (Linzen, 2016). Hence, vectors for the analogy words may only exhibit local consistency, with Eq. (2) breaking down for relatively distant regions. This caused the locality of linearity that has been reported by Nakashole & Flauger (2018), Li et al. (2021) and Wang et al. (2021a).

The second reason why a CLWE mapping may not be linear is semantic gaps. Not all analogies are language-agnostic. For example, languages pairs may have very different grammars, e.g., Chinese does have the plural morphology, so some types of analogy, e.g. `G-PL` used above, do not hold. Also, analogies may evolve differently across cultures, (see example in the lower part of Fig. 3). These two factors go some way to explain why typologically and etymologically distant language pairs tend to have worse alignment (Ruder et al., 2019). In addition, many studies point out that differences in the domain of training data can influence the similarity between multilingual word embeddings (Søgaard et al., 2018; Artetxe et al., 2018). Besides, we argue that due to polysemy, analogies may change from one domain to another, also violating Eq. (2).

Mitigating Non-Linearity. The proposed analogy-inspired framework justifies the success and failure of the linearity assumption for CLWEs. As discussed earlier, it also suggests a method for indirectly assessing the linearity of a CLWE mapping prior to implementation. Moreover, it offers principled methods for designing more effective CLWE methods. The most straightforward idea is to explicitly use Eq. (2) as a training constraint, which has very recently been practised by Anonymous (2021)[10]. Based on analogy pairs retrieved from external knowledge bases for different languages, their approach directly learnt to better encode monolingual analogies, particularly those whose vectors are distant in the embedding space. It not only works well on static word embeddings, but also leads to performance gain for large-scale pre-trained cross-lingual language models including the multilingual BERT (Devlin et al., 2019). These results on multiple tasks (e.g., bilingual lexicon induction and cross-lingual sentence retrieval) can be seen as an independent confirmation of this paper's main claim and demonstration of its usefulness.

Our study also suggests another unexplored direction: incorporating analogy-based information into non-linear CLWE mappings. Existing work has already introduced non-linearity to CLWE mappings by applying a variety of techniques including directly training non-linear functions (Mohiuddin et al., 2020), tuning linear mappings for outstanding non-isomorphic instances (Glavaš & Vulić, 2020) and learning multiple linear CLWE mappings instead of a single one (Nakashole, 2018; Wang et al., 2021a) (see § 2). However, there is a lack of theoretical motivation for decisions about how the non-linear mapping should be modelled. Nevertheless, the results presented here suggest that ensembles of linear transformations, covering analogy preserving regions of the embedding space, would make a reasonable approximation of the ground-truth CLWE mappings

and that information about analogy preservation could be used to partition embedding spaces into multiple regions, between which independent linear mappings can be learnt. We leave this application as our important future work.

7 Conclusion and Future Work

This paper makes the first attempt to explore the conditions under which CLWE mappings are linear. Theoretically, we show that this widely-adopted assumption holds *iff* the analogies encoded are preserved across embeddings for different languages. We describe the construction of a novel cross-lingual word analogy dataset for a diverse range of languages and analogy categories and we propose indicators to quantify linearity and analogy preservation. Experiment results on three distinct embedding series firmly support our hypothesis. We also demonstrate how our insight into the connection between linearity and analogy preservation can be used to better understand past observations about the limitations of linear CLWE mappings, particularly when they are ineffective. Our findings regarding the preservation of analogy encoding provide a test that can be applied to determine the likely success of any attempt to create linear mappings between multilingual embeddings. We hope this study can guide future studies in the CLWE field.

Additionally, we plan to expand our theoretical insight to contextual embeddings, inspired by Anonymous (2021) who demonstrated that developing mappings that preserve encoded analogies benefits pre-trained cross-lingual language models as well. We also aim to enrich xANLG by including new languages and analogies to enable explorations at an even larger scale. Finally, we will further design CLWE approaches

[10]Anonymous (2021) cited our earlier preprint as the primary motivation for their approach. To respect the double-blind reviewing system, we have to anonymise their reference entry here and in § 7.

that learn multiple linear mappings between local embedding regions outlined with analogy-based metrics (see § 6).

Broader Impact Statement

CLWE bridges the gap between languages and is efficient enough to be applied in situations where limited resources are available, including to endangered languages (Zhang et al., 2020; Ngoc Le & Sadat, 2020). This paper presented a theoretical analysis of the mechanisms underlying CLWE techniques which has potential to improve these methods. Our analysis relies on the use of analogies and previous work has indicated that these may contain biases, e.g., regarding gender (Bolukbasi et al., 2016; Sun et al., 2019). Any future work that incorporates analogies within the CLWE process should be aware of the potential consequences of any biases that may be contained within the analogies used. On the other hand, there is potential for the findings of this work to be leveraged for bias alleviation in cross-lingual representation learning.

References

Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 223–231, Long Beach, California, USA, 2019. PMLR. URL `http://proceedings.mlr.press/v97/allen19a.html`.

Anonymous. Will be revealed after the double-blind review. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016. doi: 10.1162/tacl_a_00106. URL `https://www.aclweb.org/anthology/Q16-1028`.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2294, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1250. URL `https://www.aclweb.org/anthology/D16-1250`.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 451–462, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1042. URL `https://www.aclweb.org/anthology/P17-1042`.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 789–798, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073. URL `https://www.aclweb.org/anthology/P18-1073`.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL `https://aclanthology.org/Q17-1010`.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Tomáš Brychcín, Stephen Taylor, and Lukáš Svoboda. Cross-lingual word analogies using linear transformations between semantic spaces. *Expert Systems with Applications*, 135:287 – 295, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.06.021. URL `http://www.sciencedirect.com/science/article/pii/S0957417419304191`.

Cristian Cardellino. Spanish Billion Words Corpus and Embeddings, August 2019. URL https://crscardellino.github.io/SBWCE/.

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*, pp. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Aleksandr Drozd, Anna Gladkova, and Satoshi Matsuoka. Word embeddings, analogies, and machine learning: Beyond king - man + woman = queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3519–3530, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1332.

Haim Dubossarsky, Ivan Vulić, Roi Reichart, and Anna Korhonen. The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2377–2390, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.186. URL https://aclanthology.org/2020.emnlp-main.186.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards understanding linear word analogies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3253–3262, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1315. URL https://www.aclweb.org/anthology/P19-1315.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 462–471, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1049. URL https://www.aclweb.org/anthology/E14-1049.

R.A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925. URL http://psychclassics.yorku.ca/Fisher/Methods/.

Louis Fournier and Ewan Dunbar. Paraphrases do not explain word analogies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2129–2134, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.182. URL https://aclanthology.org/2021.eacl-main.182.

Ashwinkumar Ganesan, Francis Ferraro, and Tim Oates. Learning a reversible embedding mapping using bi-directional manifold alignment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3132–3139, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.276. URL https://aclanthology.org/2021.findings-acl.276.

Nicolas Garneau, Mareike Hartmann, Anders Sandholm, Sebastian Ruder, Ivan Vulić, and Anders Søgaard. Analogy training multilingual encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (14):12884–12892, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17524.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(83)80009-3. URL https://www.sciencedirect.com/science/article/pii/S0364021383800093.

Alex Gittens, Dimitris Achlioptas, and Michael W. Mahoney. Skip-Gram - Zipf + Uniform = Vector Additivity. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 69–76, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1007. URL https://www.aclweb.org/anthology/P17-1007.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 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*, pp. 8–15, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2002. URL https://aclanthology.org/N16-2002.

Goran Glavaš and Ivan Vulić. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7548–7555, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.675. URL https://aclanthology.org/2020.acl-main.675.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1550.

Tatsunori B. Hashimoto, David Alvarez-Melis, and Tommi S. Jaakkola. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286, 2016. doi: 10.1162/tacl_a_00098. URL https://aclanthology.org/Q16-1020.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. Data filtering using cross-lingual word embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 162–172, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.15. URL https://aclanthology.org/2021.naacl-main.15.

O. Kariv and S. L. Hakimi. An algorithmic approach to network location problems. II: The P-Medians. *SIAM Journal on Applied Mathematics*, 37(3):539–560, 1979. doi: 10.1137/0137041. URL https://doi.org/10.1137/0137041.

Yunsu Kim, Miguel Graça, and Hermann Ney. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 35–44, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://www.aclweb.org/anthology/2020.eamt-1.5.

Martin Kleiber and W. J. Pervin. A generalized Banach-Mazur theorem. *Bulletin of the Australian Mathematical Society*, 1(2):169–173, 1969. doi: 10.1017/S0004972700041411.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th International Conference on Computational Semantics*, pp. 40–45, London, UK, April 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W15-0105.

Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018a. URL https://openreview.net/forum?id=rkYTTf-AZ.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018b. URL https://openreview.net/forum?id=H196sainb.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 171–180, Ann Arbor, Michigan, June 2014a. Association for Computational Linguistics. doi: 10.3115/v1/W14-1618. URL https://www.aclweb.org/anthology/W14-1618.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014b. URL https://proceedings.neurips.cc/paper/2014/file/feab05aa91085b7a8012516bc3533958-Paper.pdf.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015. doi: 10.1162/tacl_a_00134. URL https://aclanthology.org/Q15-1016.

Yuling Li, Kui Yu, and Yuhong Zhang. Learning cross-lingual mappings in imperfectly isomorphic embedding spaces. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2630–2642, 2021. doi: 10.1109/TASLP.2021.3097935.

Tal Linzen. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 13–18, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2503. URL https://www.aclweb.org/anthology/W16-2503.

Noa Yehezkel Lubin, Jacob Goldberger, and Yoav Goldberg. Aligning vector-spaces with noisy supervised lexicon. 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)*, pp. 460–465, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1045. URL https://www.aclweb.org/anthology/N19-1045.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013a. URL https://openreview.net/forum?id=idpCdOWtqXd60.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013b. URL http://arxiv.org/abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 3111–3119, USA, 2013c. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999959.

Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. LNMap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2712–2723, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.215. URL https://aclanthology.org/2020.emnlp-main.215.

Ndapa Nakashole. NORMA: Neighborhood sensitive maps for multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 512–522, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1047. URL https://www.aclweb.org/anthology/D18-1047.

Ndapa Nakashole and Raphael Flauger. Characterizing departures from linearity in word translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 221–227, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2036. URL https://www.aclweb.org/anthology/P18-2036.

J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 01 1965. ISSN 0010-4620. doi: 10.1093/comjnl/7.4.308. URL https://doi.org/10.1093/comjnl/7.4.308.

Tan Ngoc Le and Fatiha Sadat. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4661–4666, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.410. URL https://aclanthology.org/2020.coling-main.410.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 184–193, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1018. URL https://www.aclweb.org/anthology/P19-1018.

Xutan Peng, Yi Zheng, Chenghua Lin, and Advaith Siddharthan. Summarising historical text in modern languages. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3123–3142, Online, April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.eacl-main.273.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

Henri Prade and Gilles Richard. Analogical proportions: Why they are useful in ai. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4568–4576. International Joint Conferences on Artificial Intelligence Organization, 8 2021. doi: 10.24963/ijcai.2021/621. URL https://doi.org/10.24963/ijcai.2021/621. Survey Track.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65(1):569–630, May 2019. ISSN 1076-9757. doi: 10.1613/jair.1.11640. URL https://doi.org/10.1613/jair.1.11640.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 778–788, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1072. URL https://www.aclweb.org/anthology/P18-1072.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. Cross-cultural similarity features for cross-lingual transfer learning of pragmatically motivated tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2403–2414, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.204. URL https://aclanthology.org/2021.eacl-main.204.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://aclanthology.org/P19-1159.

Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. Multilingual culture-independent word analogy datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4074–4080, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.501.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3178–3192, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.257. URL https://aclanthology.org/2020.emnlp-main.257.

Haozhou Wang, James Henderson, and Paola Merlo. Multi-adversarial learning for cross-lingual word embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 463–472, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.39. URL https://aclanthology.org/2021.naacl-main.39.

Meihong Wang, Linling Qiu, and Xiaoli Wang. A survey on knowledge graph embeddings for link prediction. *Symmetry*, 13(3), 2021b. ISSN 2073-8994. doi: 10.3390/sym13030485. URL https://www.mdpi.com/2073-8994/13/3/485.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. Cross-lingual alignment vs joint training: A comparative study and A simple unified framework. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=S1l-C0NtwS.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37 – 52, 1987. ISSN 0169-7439. doi: https://doi.org/10.1016/0169-7439(87)80084-9. URL http://www.sciencedirect.com/science/article/pii/0169743987800849. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1006–1011, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1104. URL https://www.aclweb.org/anthology/N15-1104.

Mozhi Zhang, Yoshinari Fujinuma, and Jordan Boyd-Graber. Exploiting cross-lingual subword similarities in low-resource document classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 9547–9554, 2020.

Yi Zhang, Jie Lu, Feng Liu, Qian Liu, Alan Porter, Hongshu Chen, and Guangquan Zhang. Does deep learning help topic extraction? a kernel k-means clustering method with word embedding. *Journal of Informetrics*, 12(4):1099–1117, 2018. ISSN 1751-1577. doi: https://doi.org/10.1016/j.joi.2018.09.004. URL https://www.sciencedirect.com/science/article/pii/S1751157718300257.

Yucheng Zhou, Xiubo Geng, Tao Shen, Wenqiang Zhang, and Daxin Jiang. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5822–5834, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.465. URL https://aclanthology.org/2021.naacl-main.465.

A Question Formulations

For an analogy category with t word pairs, $\binom{t}{2}$ four-item elements can be composed. An arbitrary element, $\alpha{:}\beta :: \gamma{:}\theta$, can yield eight analogy completion questions as follows:

$$\alpha{:}\beta :: \gamma{:}? \quad \beta{:}\alpha :: \theta{:}? \quad \gamma{:}\alpha :: \theta{:}? \quad \theta{:}\beta :: \gamma{:}?$$
$$\alpha{:}\gamma :: \beta{:}? \quad \beta{:}\theta :: \alpha{:}? \quad \gamma{:}\theta :: \alpha{:}? \quad \theta{:}\gamma :: \beta{:}?$$

Hence, $\binom{t}{2} \times 8$ unique questions can be generated.

B Raw Data for Tab. 2

xANLG$_{\mathrm{G}}$		EN-DE	EN-ES	EN-FR	EN-HI	EN-PL	DE-ES	DE-FR	DE-HI	DE-PL	ES-FR	ES-HI	ES-PL	FR-HI	FR-PL	HI-PL
Wiki	CAP	.16	.21	.17	.36	.23	.21	.18	.36	.22	.22	.35	.25	.35	.23	.33
	GNDR	.32	.42	.39	.26	.35	.48	.40	.41	.36	.39	.43	.38	.30	.40	.42
	NATL	.18	.16	.15	.14	.20	.19	.19	.33	.21	.16	.30	.21	.14	.20	.32
	G-PL	.22	.23	.22	.36	.26	.25	.23	.35	.26	.25	.38	.27	.37	.26	.38
Crawl	CAP	.23	.23	.20	.23	.29	.26	.23	.24	.28	.23	.26	.28	.24	.29	.38
	GNDR	.57	.58	.59	.56	.54	.65	.66	.57	.59	.64	.56	.57	.56	.57	.58
	NATL	.32	.43	.27	.39	.29	.32	.35	.47	.35	.40	.43	.31	.46	.31	.42
	G-PL	.35	.24	.33	.48	.29	.33	.37	.44	.42	.33	.47	.33	.48	.42	.51
CoNLL	CAP	.31	.58	.32	.55	.39	.58	.32	.56	.38	.59	.66	.59	.56	.40	.55
	GNDR	.48	.76	.49	.55	.48	.74	.55	.57	.50	.77	.76	.72	.59	.52	.58
	NATL	.37	.72	.26	.51	.38	.78	.34	.52	.36	.74	.74	.73	.50	.35	.50
	G-PL	.32	.67	.32	.48	.36	.65	.34	.47	.36	.68	.67	.65	.50	.38	.49

xANLG$_{\mathrm{M}}$		EN ET	EN FI	EN HR	EN LV	EN RU	EN SL	ET FI	ET HR	ET LV	ET RU	ET SL	FI HR	FI LV	FI RU	FI SL	HR LV	HR RU	HR SL	LV RU	LV SL	RU SL
Wiki	ANIM	.50	.50	.22	.31	.19	.15	.56	.27	.37	.30	.35	.29	.41	.30	.40	.32	.36	.28	.31	.22	.20
	G-PL	.25	.22	.37	.37	.28	.33	.24	.31	.29	.28	.26	.30	.29	.26	.27	.33	.32	.30	.33	.28	.28
Crawl	ANIM	.55	.55	.55	.49	.55	.51	.34	.41	.45	.22	.41	.40	.46	.41	.45	.37	.23	.28	.38	.24	.43
	G-PL	.28	.43	.47	.43	.45	.40	.30	.45	.37	.43	.37	.46	.40	.44	.43	.42	.50	.54	.39	.35	.43
CoNLL	ANIM	.54	.54	.99	.55	.50	.53	.29	.74	.46	.37	.43	.87	.51	.38	.46	.64	.77	.98	.42	.36	.41
	G-PL	.45	.40	.52	.42	.40	.42	.37	.77	.41	.41	.40	.81	.37	.36	.39	.84	.66	.77	.36	.40	.38

Table 3: Raw $\mathcal{S}_{\mathrm{LMP}}$ results (the negative sign is omitted for brevity).

	Wiki				Crawl				CoNLL			
	CAP	GNDR	NATL	G-PL	CAP	GNDR	NATL	G-PL	CAP	GNDR	NATL	G-PL
DE	.68	.25	.21	.23	.47	.48	.79	.77	.65	.43	.41	.55
EN	.94	.33	.94	.58	.57	.67	.76	.94	.87	.57	.79	.61
ES	.45	.13	.35	.13	.40	.57	.68	.87	.13	.07	.07	.17
FR	.92	.27	.76	.13	.65	.50	.85	.87	.48	.14	.24	.35
HI	.29	.30	.42	.07	.58	.59	.59	.32	.32	.37	.31	.16
PL	.16	.21	.26	.10	.29	.55	.82	.84	.45	.45	.38	.52

	Wiki		Crawl		CoNLL	
	ANIM	G-PL	ANIM	G-PL	ANIM	G-PL
EN	.48	.65	.29	.87	.36	.58
ET	.12	.50	.52	1.00	.21	.48
FI	.06	.65	.48	.87	.42	.54
HR	.17	.20	.50	.68	.07	.11
LV	.19	.10	.39	.84	.27	.23
RU	.36	.40	.61	.87	.42	.55
SL	.42	.23	.39	.81	.12	.39

Table 4: Raw monolingual LRCos results (left:xANLG$_{\mathrm{G}}$; right: xANLG$_{\mathrm{M}}$).