Multilingual Factor Analysis

Francisco Vargas, Kamen Brestnichki, Alex Papadopoulos-Korfiatis and **Nils Hammerla**
Babylon Health
{firstname.lastname, alex.papadopoulos}@babylonhealth.com

Abstract

In this work we approach the task of learning multilingual word representations in an offline manner by fitting a generative latent variable model to a multilingual dictionary. We model equivalent words in different languages as different views of the same word generated by a common latent variable representing their latent lexical meaning. We explore the task of alignment by querying the fitted model for multilingual embeddings achieving competitive results across a variety of tasks. The proposed model is robust to noise in the embedding space making it a suitable method for distributed representations learned from noisy corpora.

1 Introduction

Popular approaches for multilingual alignment of word embeddings base themselves on the observation in (Mikolov et al., 2013a), which noticed that continuous word embedding spaces (Mikolov et al., 2013b; Pennington et al., 2014; Bojanowski et al., 2017; Joulin et al., 2017) exhibit similar structures across languages. This observation has led to multiple successful methods in which a direct linear mapping between the two spaces is learned through a least squares based objective (Mikolov et al., 2013a; Smith et al., 2017; Xing et al., 2015) using a paired bilingual dictionary.

An alternate set of approaches based on Canonical Correlation Analysis (CCA) (Knapp, 1978) seek to project monolingual embeddings into a shared multilingual space (Faruqui and Dyer, 2014b; Lu et al., 2015). Both these methods aim to exploit the correlations between the monolingual vector spaces when projecting into the aligned multilingual space. The multilingual embeddings from (Faruqui and Dyer, 2014b; Lu et al., 2015) are shown to improve on word level semantic tasks, which sustains the authors' claim that multilingual information enhances semantic spaces.

In this paper we present a new non-iterative method based on variants of factor analysis (Browne, 1979; McDonald, 1970; Browne, 1980) for aligning monolingual representations into a multilingual space. Our generative modelling assumes that a single word translation pair is generated by an embedding representing the lexical meaning of the underlying concept. We achieve competitive results across a wide range of tasks compared to state-of-the-art methods, and we conjecture that our multilingual latent variable model has sound generative properties that match those of psycholinguistic theories of the bilingual mind (Weinreich, 1953). Furthermore, we show how our model extends to more than two languages within the generative framework which is something that previous alignment models are not naturally suited to, instead resorting to combining bilingual models with a pivot as in (Ammar et al., 2016).

Additionally the general benefit of the probabilistic setup as discussed in (Tipping and Bishop, 1999) is that it offers the potential to extend the scope of conventional alignment methods to model and exploit linguistic structure more accurately. An example of such a benefit could be modelling how corresponding word translations can be generated by more than just a single latent concept. This assumption can be encoded by a mixture of Factor Analysers (Ghahramani et al., 1996) to model word polysemy in a similar fashion to (Athiwaratkun and Wilson, 2017), where mixtures of Gaussians are used to reflect the different meanings of a word.

The main contribution of this work is the application of a well-studied graphical model to a novel domain, outperforming previous approaches on word and sentence-level translation retrieval

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1738–1750
Florence, Italy, July 28 - August 2, 2019.

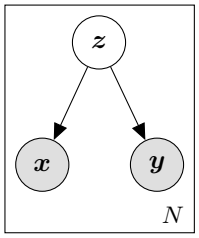

Figure 1: Graphical model for alignment. Latent space $\boldsymbol{z}$ represents the aligned shared space between the two vector spaces $\boldsymbol{x}$ and $\boldsymbol{y}$.

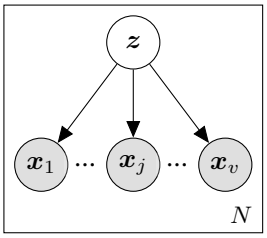

Figure 2: Graphical model for MBFA. Latent space $\boldsymbol{z}$ represents the aligned shared space between the multiple vector spaces $\{\boldsymbol{x}_j\}_{j=1}^{v}$.

tasks. We put the model through a battery of tests, showing it aligns embeddings across languages well, while retaining performance on monolingual word-level and sentence-level tasks. Finally, we apply a natural extension of this model to more languages in order to align three languages into a single common space.

2 Background

Previous work on the topic of embedding alignment has assumed that alignment is a directed procedure — i.e. we want to align French to English embeddings. However, another approach would be to align both to a common latent space that is not necessarily the same as either of the original spaces. This motivates applying a well-studied latent variable model to this problem.

2.1 Factor Analysis

Factor analysis (Spearman, 1904; Thurstone, 1931) is a technique originally developed in psychology to study the correlation of latent factors $\boldsymbol{z} \in \mathbb{R}^k$ on observed measurements $\boldsymbol{x} \in \mathbb{R}^d$. Formally:

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbb{I}),$$
$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \boldsymbol{\Psi}).$$

In order to learn the parameters $\boldsymbol{W}, \boldsymbol{\Psi}$ of the model we maximise the marginal likelihood $p(\boldsymbol{x}|\boldsymbol{W}, \boldsymbol{\Psi})$ with respect to $\boldsymbol{W}, \boldsymbol{\Psi}$. The maximum likelihood estimates of these procedures can be used to obtain latent representations for a given observation $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\boldsymbol{z}]$. Such projections have been found to be generalisations of principal component analysis (Pearson, 1901) as studied in (Tipping and Bishop, 1999).

2.2 Inter-Battery Factor Analysis

Inter-Battery Factor Analysis (IBFA) (Tucker, 1958; Browne, 1979) is an extension of factor analysis that adapts it to two sets of variables $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{y} \in \mathbb{R}^{d'}$ (i.e. embeddings of two languages). In this setting it is assumed that pairs of observations are generated by a shared latent variable $\boldsymbol{z}$

$$\begin{aligned} p(\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{z}; \mathbf{0}, \mathbb{I}), \\ p(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{x}; \boldsymbol{W}_x\boldsymbol{z} + \boldsymbol{\mu}_x, \boldsymbol{\Psi}_x), \\ p(\boldsymbol{y}|\boldsymbol{z}) &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{W}_y\boldsymbol{z} + \boldsymbol{\mu}_y, \boldsymbol{\Psi}_y). \end{aligned} \quad (1)$$

As in traditional factor analysis, we seek to estimate the parameters that maximise the marginal likelihood

$$\begin{aligned} &\arg\max_{\{\boldsymbol{\Psi}_i, \boldsymbol{W}_i\}} \prod_k p(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)} | \{\boldsymbol{\Psi}_i, \boldsymbol{W}_i\}_i), \\ &\text{subject to } \boldsymbol{\Psi}_i \succ \mathbf{0},\ (\boldsymbol{W}_i^\top \boldsymbol{W}_i) \succcurlyeq \mathbf{0}, \end{aligned} \quad (2)$$

where the joint marginal $p(\boldsymbol{x}_k, \boldsymbol{y}_k | \{\boldsymbol{\Psi}_i, \boldsymbol{W}_i\}_i)$ is a Gaussian with the form

$$\mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_{\boldsymbol{x}}\\ \boldsymbol{\mu}_{\boldsymbol{y}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right),$$
$$\boldsymbol{\Sigma}_{ij} = \boldsymbol{W}_i\boldsymbol{W}_j^\top + \delta_{ij}\boldsymbol{\Psi}_i,$$

and $\boldsymbol{\Psi} \succ \mathbf{0}$ means $\boldsymbol{\Psi}$ is positive definite.

Maximising the likelihood as in Equation 2 will find the optimal parameters for the generative process described in Figure 1 where one latent $\boldsymbol{z}$ is responsible for generating a pair $\boldsymbol{x}, \boldsymbol{y}$. This makes it a suitable objective for aligning the vector spaces of $\boldsymbol{x}$, $\boldsymbol{y}$ in the latent space. In contrast to the discriminative directed methods in (Mikolov et al., 2013a; Smith et al., 2017; Xing et al., 2015), IBFA has the capacity to model noise.

We can re-interpret the logarithm of Equation 2

(as shown in Appendix D) as

$$\sum_k \log p(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)}|\boldsymbol{\theta}) = C + \sum_k (\mathcal{L}_k^{y|x} + \mathcal{L}_k^x), \quad (3)$$

$$\mathcal{L}_k^{y|x} = -\frac{1}{2}\|\tilde{\boldsymbol{y}}^{(k)} - \boldsymbol{W}_y \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x}^{(k)})}[\boldsymbol{z}]\|^2_{\boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}}},$$

$$\mathcal{L}_k^{x} = -\frac{1}{2}\|\tilde{\boldsymbol{x}}^{(k)} - \boldsymbol{W}_x \mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x}^{(k)})}[\boldsymbol{z}]\|^2_{\boldsymbol{\Psi}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Psi}_{\boldsymbol{x}}},$$

$$C = -\frac{N}{2}(\log|2\pi\boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}}| + \log|2\pi\boldsymbol{\Sigma}_{\boldsymbol{x}}|).$$

The exact expression for $\boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}}$ is given in the same appendix. This interpretation shows that for each pair of points, the objective is to minimise the reconstruction errors of $\boldsymbol{x}$ and $\boldsymbol{y}$, given a projection into the latent space $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x}_k)}[\boldsymbol{z}]$. By utilising the symmetry of Equation 2, we can show the converse is true as well — maximising the joint probability also minimises the reconstruction loss given the latent projections $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{y}_k)}[\boldsymbol{z}]$. Thus, this forces the latent embeddings of $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ to be close in the latent space. This provides intuition as to why embedding into this common latent space is a good alignment procedure.

In (Browne, 1979; Bach and Jordan, 2005) it is shown that the maximum likelihood estimates for $\{\boldsymbol{\Psi}_i, \boldsymbol{W}_i\}$ can be attained in closed form

$$\hat{\boldsymbol{W}}_i = \boldsymbol{S}_{ii}\boldsymbol{U}_i\boldsymbol{P}^{1/2},$$
$$\hat{\boldsymbol{\Psi}}_i = \boldsymbol{S}_{ii} - \hat{\boldsymbol{W}}_i\hat{\boldsymbol{W}}_i^\top,$$
$$\hat{\boldsymbol{\mu}}_x = \bar{\boldsymbol{x}},\ \hat{\boldsymbol{\mu}}_y = \bar{\boldsymbol{y}},$$

where

$$\boldsymbol{S}_{xx} = \frac{1}{m}\sum_{i=1}^m \tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top},$$
$$\boldsymbol{S}_{yy} = \frac{1}{m}\sum_{i=1}^m \tilde{\boldsymbol{y}}^{(i)}\tilde{\boldsymbol{y}}^{(i)\top},$$
$$\boldsymbol{U}_i = \boldsymbol{S}_{ii}^{-1/2}\boldsymbol{V}_i,$$
$$\boldsymbol{V}_x\boldsymbol{P}\boldsymbol{V}_y^\top = \text{SVD}(\boldsymbol{S}_{xx}^{-1/2}\boldsymbol{S}_{xy}\boldsymbol{S}_{yy}^{-1/2}).$$

The projections into the latent space from $\boldsymbol{x}$ are given by (as proved in Appendix B)

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\boldsymbol{z}] = (\mathbb{I} + \boldsymbol{W}_x^\top\boldsymbol{\Psi}_x^{-1}\boldsymbol{W}_x)^{-1}\boldsymbol{W}_x^\top\boldsymbol{\Psi}_x^{-1}\tilde{\boldsymbol{x}},$$
$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{\mu}_x. \quad (4)$$

Evaluated at the MLE, (Bach and Jordan, 2005) show that Equation 4 can be reduced to

$$\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\boldsymbol{z}] = \boldsymbol{P}^{1/2}\boldsymbol{U}_x^\top(\boldsymbol{x} - \boldsymbol{\mu}_x).$$

2.2.1 Multiple-Battery Factor Analysis

Multiple-Battery Factor Analysis (MBFA) (McDonald, 1970; Browne, 1980) is a natural extension of IBFA that models more than two views of observables (i.e. multiple languages), as shown in Figure 2.

Formally, for a set of views $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_v\}$, we can write the model as

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \mathbb{I}),$$
$$p(\boldsymbol{x}_i|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{W}_i\boldsymbol{z} + \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i).$$

Similar to IBFA the projections to the latent space are given by Equation 4, and the marginal yields a very similar form

$$\mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}_1\\ \vdots \\ \boldsymbol{x}_v\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_1\\ \vdots \\ \boldsymbol{\mu}_v\end{bmatrix}, \begin{bmatrix}\boldsymbol{W}_1\boldsymbol{W}_1^\top + \boldsymbol{\Psi}_1 \dots & \boldsymbol{W}_1\boldsymbol{W}_v^\top \\ \vdots & \ddots & \vdots \\ \boldsymbol{W}_v\boldsymbol{W}_1^\top & \dots \boldsymbol{W}_v\boldsymbol{W}_v^\top + \boldsymbol{\Psi}_v\end{bmatrix}\right).$$

Unlike IBFA, a closed form solution for maximising the marginal likelihood of MBFA is unknown. Because of this, we have to resort to iterative approaches as in (Browne, 1980) such as the natural extension of the EM algorithm proposed by (Bach and Jordan, 2005). Defining

$$\boldsymbol{M}_t = \left(\mathbb{I} + \boldsymbol{W}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{W}_t\right)^{-1},$$
$$\boldsymbol{B}_t = \boldsymbol{M}_t\boldsymbol{W}_t^\top\boldsymbol{\Psi}_t^{-1},$$
$$\widetilde{\boldsymbol{\Psi}}_{t+1} = \boldsymbol{S} - \boldsymbol{S}\boldsymbol{\Psi}_t^{-1}\boldsymbol{W}_t\boldsymbol{M}_t^\top\boldsymbol{W}_{t+1}^\top,$$

the EM updates are given by

$$\boldsymbol{W}_{t+1} = \boldsymbol{S}\boldsymbol{B}_t^\top\left(\boldsymbol{M}_t + \boldsymbol{B}_t\boldsymbol{S}\boldsymbol{B}_t^\top\right)^{-1},$$
$$\boldsymbol{\Psi}_{t+1} = \text{Bdiag}\left((\widetilde{\boldsymbol{\Psi}}_{t+1})_{11}, \dots, (\widetilde{\boldsymbol{\Psi}}_{t+1})_{vv}\right),$$

where $\boldsymbol{S}$ is the sample covariance matrix of the concatenated views (derivation provided in Appendix E). (Browne, 1980) shows that, under suitable conditions, the MLE of the parameters of MBFA is uniquely identifiable (up to a rotation that does not affect the method's performance). We observed this in an empirical study — the solutions we converge to are always a rotation away from each other, irrespective of the parameters' initialisation. This heavily suggests that any optimum is a global optimum and thus we restrict ourselves to only reporting results we observed when fitting from a single initialisation. The chosen initialisation point is provided as Equation (3.25) of (Browne, 1980).

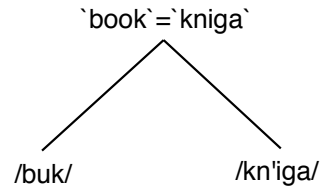

Figure 3: Weinrich's compound model for lexical association between English and Russian. Image from (Neuser, 2017).

3 Multilingual Factor Analysis

We coin the term Multilingual Factor Analysis for the application of methods based on IBFA and MBFA to model the generation of multilingual tuples from a shared latent space. We motivate our generative process with the compound model for language association presented by (Weinreich, 1953). In this model a lexical meaning entity (a concept) is responsible for associating the corresponding words in the two different languages.

We note that the structure in Figure 3 is very similar to our graphical model for IBFA specified in Figure 1. We can interpret our latent variable as the latent lexical concept responsible for associating (generating) the multilingual language pairs. Most theories that explain the interconnections between languages in the bilingual mind assume that "while phonological and morphosyntactic forms differ across languages, meanings and/or concepts are largely, if not completely, shared" (Pavlenko, 2009). This shows that our generative modelling is supported by established models of language interconnectedness in the bilingual mind.

Intuitively, our approach can be summarised as transforming monolingual representations by mapping them to a concept space in which lexical meaning across languages is aligned and then performing retrieval, translation and similarity-based tasks in that aligned concept space.

3.1 Comparison to Direct Methods

Methods that learn a direct linear transformation from $\boldsymbol{x}$ to $\boldsymbol{y}$, such as (Mikolov et al., 2013a; Artetxe et al., 2016; Smith et al., 2017; Lample et al., 2018) could also be interpreted as maximising the conditional likelihood

$$\prod_k p(\boldsymbol{y}^{(k)}|\boldsymbol{x}^{(k)}) = \prod_k \mathcal{N}(\boldsymbol{y}^{(k)}; \boldsymbol{W}\boldsymbol{x}^{(k)} + \boldsymbol{\mu}, \boldsymbol{\Psi}).$$

As shown in Appendix F, the maximum likelihood estimate for $\boldsymbol{W}$ does not depend on the noise term $\boldsymbol{\Psi}$. In addition, even if one were to fit $\boldsymbol{\Psi}$, it is not clear how to utilise it to make predictions as the conditional expectation

$$\mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{x}^{(k)})}[\boldsymbol{y}] = \boldsymbol{W}\boldsymbol{x}^{(k)} + \boldsymbol{\mu},$$

does not depend on the noise parameters. As this method is therefore not robust to noise, previous work has used extensive regularisation (i.e. by making $\boldsymbol{W}$ orthogonal) to avoid overfitting.

3.2 Relation to CCA

CCA is a popular method used for multilingual alignment which is very closely related to IBFA, as detailed in (Bach and Jordan, 2005). (Barber, 2012) shows that CCA can be recovered as a limiting case of IBFA with constrained diagonal covariance $\boldsymbol{\Psi}_x = \sigma_x^2\mathbb{I}$, $\boldsymbol{\Psi}_y = \sigma_y^2\mathbb{I}$, as $\sigma_x^2, \sigma_y^2 \to 0$. CCA assumes that the emissions from the latent spaces to the observables are deterministic. This is a strong and unrealistic assumption given that word embeddings are learned from noisy corpora and stochastic learning algorithms.

4 Experiments

In this section, we empirically demonstrate the effectiveness of our generative approach on several benchmarks, and compare it with state-of-the-art methods. We first present cross-lingual (word-translation) evaluation tasks to evaluate the quality of our multi-lingual word embeddings. As a follow-up to the word retrieval task we also run experiments on cross-lingual sentence retrieval tasks. We further demonstrate the quality of our multi-lingual word embeddings on monolingual word- and sentence-level similarity tasks from (Faruqui and Dyer, 2014b), which we believe provides empirical evidence that the aligned embeddings preserve and even potentially enhance their monolingual quality.

4.1 Word Translation

This task is concerned with the problem of retrieving the translation of a given set of source words. We reproduce results in the same environment as (Lample et al., 2018)[1] for a fair comparison. We perform an ablation study to assess the effectiveness of our method in the Italian to English (it-en) setting in (Smith et al., 2017; Dinu et al., 2014).

[1] github.com/Babylonpartners/MultilingualFactorAnalysis, based on github.com/facebookresearch/MUSE.

Method	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en
Supervised										
SVD	77.4	77.3	74.9	76.1	68.4	67.7	**47.0**	58.2	27.3*	09.3*
IBFA	**79.5**	**81.5**	**77.3**	**79.5**	**70.7**	**72.1**	46.7	**61.3**	**42.9**	**36.9**
SVD+CSLS	81.4	82.9	81.1	82.4	73.5	72.4	**51.7**	63.7	32.5*	25.1*
IBFA+CSLS	**81.7**	**84.1**	**81.9**	**83.4**	**74.1**	**75.7**	50.5	**66.3**	**48.4**	**41.7**
Semi-supervised										
SVD	65.9	74.1	71.0	72.7	60.3	65.3	11.4	37.7	06.8	00.8
IBFA	76.1	**80.1**	77.1	**78.9**	66.8	**71.8**	23.1	39.9	17.1	**24.0**
AdvR	**79.1**	78.1	**78.1**	78.2	**71.3**	69.6	**37.3**	**54.3**	**30.9**	21.9
SVD+CSLS	73.0	80.7	75.7	79.6	65.3	70.8	20.9	41.5	10.5	01.7
IBFA+CSLS	76.5	**83.7**	78.6	**82.3**	68.7	**73.7**	25.3	46.3	22.1	27.2
AdvR+CSLS	**81.7**	83.3	**82.3**	82.1	**74.0**	72.2	**44.0**	**59.1**	**32.5**	**31.4**

Table 1: Precision @1 for cross-lingual word similarity tasks. Rows labelled AdvR are copies of Adversarial - Refine rows in (Lample et al., 2018). Results marked with a * differ from the ones shown in (Lample et al., 2018) due to pre-processing done on their part. SVD and IBFA in the semi-supervised setting use the pseudo-dictionary, while AdvR uses frequency information. CSLS is the post-processing technique proposed in (Lample et al., 2018).

In these experiments we are interested in studying the effectiveness of our method compared to that of the Procrustes-based fitting used in (Smith et al., 2017) without any post-processing steps to address the hubness problem (Dinu et al., 2014). In Table 1 we observe how our model is competitive to the results in (Lample et al., 2018) and outperforms them in most cases. We notice that given an expert dictionary, our method performs the best out of all compared methods on all tasks, except in English to Russian (en-ru) translation where it remains competitive. What is surprising is that, in the semi-supervised setting, IBFA bridges the gap between the method proposed in (Lample et al., 2018) on languages where the dictionary of identical tokens across languages (i.e. the pseudo-dictionary from (Smith et al., 2017)) is richer. However, even though it significantly outperforms SVD using the pseudo-dictionary, it cannot match the performance of the adversarial approach for more distant languages like English and Chinese (en-zh).

4.1.1 Detailed Comparison to Basic SVD

We present a more detailed comparison to the SVD method described in (Smith et al., 2017). We focus on methods in their base form, that is without post-processing techniques, i.e. cross-domain similarity local scaling (CSLS) (Lample et al., 2018) or inverted softmax (ISF) (Smith et al., 2017). Note that (Smith et al., 2017) used the scikit-learn [2] implementation of CCA, which uses an iterative estimation of partial least squares. This does not give the same results as the standard CCA procedure. In Table 2 we reproduce the results from (Smith et al., 2017) using the dictionaries and embeddings provided by (Dinu et al., 2014)[3] and we compare our method (IBFA) using both the expert dictionaries from (Dinu et al., 2014) and the pseudo-dictionaries as constructed in (Smith et al., 2017). We significantly outperform both SVD and CCA, especially when using the pseudo-dictionaries.

4.2 Word Similarity Tasks

This task assesses the monolingual quality of word embeddings. In this experiment, we fit both considered methods (CCA and IBFA) on the entire available dictionary of around 100k word pairs. We compare to CCA as used in (Faruqui and Dyer, 2014b) and standard monolingual word embeddings on the available tasks from (Faruqui and Dyer, 2014b). We evaluate our multilingual embeddings on the following tasks: **WS353** (Finkelstein et al., 2002); **WS-SIM**, **WS-REL** (Agirre et al., 2009); **RG65** (Rubenstein and Goodenough, 1965); **MC-30** (Miller and Charles, 1991); **MT-**

[2] A commonly used Python library for scientific computing, found at (Pedregosa et al., 2011).

[3] http://clic.cimec.unitn.it/ georgiana.dinu/down/

	English to Italian			Italian to English			English to Italian			Italian to English		
	@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
Mikolov et. al.	33.8	48.3	53.9	24.9	41.0	47.4	1.0	2.8	3.9	2.5	6.4	9.1
CCA (Sklearn)	36.1	52.7	58.1	31.0	49.9	57.0	29.1	46.4	53.0	27.0	47.0	52.3
CCA	30.9	48.1	52.7	27.7	45.5	51.0	26.5	42.5	48.1	22.8	40.1	45.5
SVD	36.9	52.7	57.9	32.2	49.6	55.7	27.1	43.4	49.3	26.2	42.1	49.0
IBFA (Ours)	**39.3**	**55.3**	**60.1**	**34.7**	**53.5**	**59.4**	**34.7**	**52.6**	**58.3**	**33.7**	**53.3**	**59.2**

Table 2: Comparisons without post-processing of methods. Results reproduced from (Smith et al., 2017) for fair comparison. **Left**: Comparisons using the same expert dictionary as (Smith et al., 2017). **Right**: Comparisons using the pseudo-dictionary from (Smith et al., 2017).

Embeddings	WS	WS-SIM	WS-REL	RG-65	MC-30	MT-287	MT-771	MEN-TR
English	73.7	78.1	68.2	79.7	81.2	67.9	66.9	76.4
IBFA en-de	**74.4**	**79.4**	**68.3**	**81.4**	**84.2**	67.2	69.4	**77.8**
IBFA en-fr	72.4	77.8	65.8	80.5	83.0	**68.2**	**69.6**	77.6
IBFA en-es	73.6	78.5	67.0	79.0	83.0	68.2	69.4	77.3
CCA en-de	71.7	76.4	64.0	76.7	82.4	63.0	64.7	75.3
CCA en-fr	70.9	76.4	63.3	76.5	81.4	63.4	65.4	74.9
CCA en-es	70.8	76.3	63.1	76.4	81.2	63.0	65.1	74.7

Table 3: Spearman correlation for English word similarity tasks. First row represents monolingual fasttext vectors (Joulin et al., 2017) in English, the rest are bilingual embeddings.

287; (Radinsky et al., 2011); **MT-771** (Halawi et al., 2012), and **MEN-TR** (Bruni et al., 2012). These tasks consist of English word pairs that have been assigned ground truth similarity scores by humans. We use the test-suite provided by (Faruqui and Dyer, 2014a)[4] to evaluate our multilingual embeddings on these datasets. This test-suite calculates similarity of words through cosine similarity in their representation spaces and then reports Spearman correlation with the ground truth similarity scores provided by humans.

As shown in Table 3, we observe a performance gain over CCA and monolingual word embeddings suggesting that we not only preserve the monolingual quality of the embeddings but also enhance it.

4.3 Monolingual Sentence Similarity Tasks

Semantic Textual Similarity (STS) is a standard benchmark used to assess sentence similarity metrics (Agirre et al., 2012, 2013, 2014, 2015, 2016). In this work, we use it to show that our alignment procedure does not degrade the quality of the embeddings at the sentence level. For both IBFA and CCA, we align English and one other language (from French, Spanish, German) using the entire dictionaries (of about 100k word pairs each) provided by (Lample et al., 2018). We then use the procedure defined in (Arora et al., 2016) to create sentence embeddings and use cosine similarity to output sentence similarity using those embeddings. The method's performance on each set of embeddings is assessed using Spearman correlation to human-produced expert similarity scores. As evidenced by the results shown in Table 4, IBFA remains competitive using any of the three languages considered, while CCA shows a performance decrease.

4.4 Crosslingual Sentence Similarity Tasks

Europarl (Koehn, 2005) is a parallel corpus of sentences taken from the proceedings of the European parliament. In this set of experiments, we focus on its English-Italian (en-it) sub-corpus, in order to compare to previous methods. We report results under the framework of (Lample et al., 2018). That is, we form sentence embeddings using the average of the tf-idf weighted word embeddings in the bag-of-words representation of the sentence. Performance is averaged over 2,000 randomly chosen source sentence queries and 200k

[4] https://github.com/mfaruqui/eval-word-vectors

Embeddings	STS12	STS13*	STS14	STS15	STS16
English	**58.1**	69.2	66.7	72.6	70.6
IBFA en-de	**58.1**	**70.2**	**66.8**	73.0	71.6
IBFA en-fr	58.0	70.0	66.7	72.8	71.4
IBFA en-es	57.9	69.7	66.6	72.9	**71.7**
CCA en-de	56.7	67.5	65.7	**73.1**	70.5
CCA en-fr	56.7	67.9	65.9	72.8	70.8
CCA en-es	56.6	67.8	65.9	72.9	70.8

Table 4: Spearman correlation for Semantic Textual Similarity (STS) tasks in English. All results use the sentence embeddings described in (Arora et al., 2016). First row represents monolingual FastText vectors (Joulin et al., 2017) in English, the rest are bilingual embeddings. *STS13 excludes the proprietary SMT dataset.

	English to Italian			Italian to English		
	@1	@5	@10	@1	@5	@10
Mikolov et. al.✓	10.5	18.7	22.8	12.0	22.1	26.7
Dinu et al.✓	45.3	72.4	80.7	48.9	71.3	78.3
Smith et al.✓	54.6	72.7	78.2	42.9	62.2	69.2
SVD	40.5	52.6	56.9	51.2	63.7	67.9
IBFA (Ours)	62.7	74.2	77.9	64.1	75.2	79.5
SVD + CSLS	64.0	75.8	78.5	67.9	79.4	82.8
AdvR + CSLS	66.2	80.4	83.4	58.7	76.5	80.9
IBFA + CSLS	**68.8**	**80.7**	**83.5**	**70.2**	**80.8**	**84.8**

Table 5: Sentence translation precisions @1, @5, @10 on 2,000 English-Italian pairs samples from a set of 200k sentences from Europarl (Koehn, 2005) on Dinu embeddings. AdvR is copied from Adversarial - Refined in (Lample et al., 2018). Rows with ✓ copied from (Smith et al., 2017).

target sentences for each language pair. Note that this is a different set up to the one presented in (Smith et al., 2017), in which an unweighted average is used. The results are reported in Table 5. As we can see, IBFA outperforms all prior methods both using nearest neighbour retrieval, where it has a gain of 20 percent absolute on SVD, as well as using the CSLS retrieval metric.

4.5 Alignment of three languages

In an ideal scenario, when we have v languages, we wouldn't want to train a transformation between each pair, as that would involve storing $\mathcal{O}(v^2)$ matrices. One way to overcome this problem is by aligning all embeddings to a common space. In this exploratory experiment, we constrain ourselves to aligning three languages at the same time, but the same methodology could be applied to an arbitrary number of languages. MBFA, the extension of IBFA described in Section 2.2.1 naturally lends itself to this task. What is needed for training this method is a dictionary of word triples across the three languages considered. We construct such a dictionary by taking the intersection of all 6 pairs of bilingual dictionaries for the three languages provided by (Lample et al., 2018). We then train MBFA for 20,000 iterations of EM (a brief analysis of convergence is provided in Appendix G). Alternatively, with direct methods like (Smith et al., 2017; Lample et al., 2018) one could align all languages to English and treat that as the common space.

We compare both approaches and present their results in Table 6. As we can see, both methods experience a decrease in overall performance when compared to models fitted on just a pair of languages, however MBFA performs better overall. That is, the direct approaches preserve their performance on translation to and from English, but translation from French to Italian decreases significantly. Meanwhile, MBFA suffers a decrease in each pair of languages, however it retains competitive performance to the direct methods on English translation. It is worth noting that as the number of aligned languages v increases, there are $O(v)$ pairs

Method	en-it	it-en	en-fr	fr-en	it-fr	fr-it
SVD	71.0	72.4	74.9	76.1	78.3	72.9
MBFA	**71.9**	**73.4**	**76.7**	**78.1**	**82.6**	**77.5**
SVD+CSLS	76.2	**77.9**	81.1	**82.4**	84.5	79.8
MBFA+CSLS	**77.4**	77.7	**81.9**	82.1	**86.8**	**81.9**

Table 6: Precision @1 when aligning English, French and Italian embeddings to a common space. For SVD, this common space is English, while for MBFA it is the latent space.

of languages, one of which is English, and $O(v^2)$ pairs in which English does not participate. This suggests that MBFA may generalise past three simultaneously aligned languages better than the direct methods.

4.6 Generating Random Word Pairs

We explore the generative process of IBFA by synthesising word pairs from noise, using a trained English-Spanish IBFA model. We follow the generative process specified in Equation 1 to generate 2,000 word vector pairs and then we find the nearest neighbour vector in each vocabulary and display the corresponding words. We then rank these 2,000 pairs according to their joint probability under the model and present the top 28 samples in Table 7. Note that whilst the sampled pairs are not exact translations, they have closely related meanings. The examples we found interesting are dreadful and despair; frightening and brutality; crazed and merry; unrealistic and questioning; misguided and conceal; reactionary and conservatism.

5 Conclusion

We have introduced a cross-lingual embedding alignment procedure based on a probabilistic latent variable model, that increases performance across various tasks compared to previous methods using both nearest neighbour retrieval, as well as the CSLS criterion. We have shown that the resulting embeddings in this aligned space preserve their quality by presenting results on tasks that assess word and sentence-level monolingual similarity correlation with human scores. The resulting embeddings also significantly increase the precision of sentence retrieval in multilingual settings. Finally, the preliminary results we have shown on aligning more than two languages at the same time provide an exciting path for future research.

en	es	es→en
particular	efectivamente	effectively
correspondingly	esto	this
silly	irónicamente	ironic
frightening	brutalidad	brutality
manipulations	intencionadamente	intentionally
ignore	contraproducente	counter-productive
fundamentally	entendido	understood
embarrassed	enojado	angry
terrified	casualidad	coincidence
hypocritical	obviamente	obviously
wondered	incómodo	uncomfortable
oftentimes	apostar	betting
unwittingly	traicionar	betray
mishap	irónicamente	ironically
veritable	empero	however
overpowered	deshacerse	fall apart
crazed	divertidos	merry
frightening	ironía	irony
dreadful	desesperación	despair
instituting	restablecimiento	recover
unrealistic	cuestionamiento	questioning
regrettable	erróneos	mistaken
irresponsible	preocupaciones	concerns
obsession	irremediablemente	hopelessly
embodied	voluntad	will
misguided	esconder	conceal
perspective	contestación	answer
reactionary	conservadurismo	conservatism

Table 7: Random pairs sampled from model, selected top 28 ranked by confidence. Proper nouns, and acronyms (names and surnames) were removed from the list. Third column represents a correct translation from Spanish to English.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * sem 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 32–43.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393. Association for Computational Linguistics.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A Smith. 2016. Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. *International Conference on Learning Representations, 2017*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Ben Athiwaratkun and Andrew Gordon Wilson. 2017. Multimodal word distributions. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, pages 427–431,Valencia, Spain, April 3-7*.

Francis R Bach and Michael I Jordan. 2005. A probabilistic interpretation of canonical correlation analysis. *Computer Science Division, University of California Berkeley*.

David Barber. 2012. *Bayesian reasoning and machine learning*, pages 474–475. Cambridge University Press.

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

Michael W Browne. 1979. The maximum-likelihood solution in inter-battery factor analysis. *British Journal of Mathematical and Statistical Psychology*, 32(1):75–86.

Michael W Browne. 1980. Factor analysis of multiple batteries by maximum likelihood. *British Journal of Mathematical and Statistical Psychology*, 33(2):184–199.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Manaal Faruqui and Chris Dyer. 2014a. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of ACL: System Demonstrations*.

Manaal Faruqui and Chris Dyer. 2014b. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131.

Zoubin Ghahramani, Geoffrey E Hinton, et al. 1996. The em algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414. ACM.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, pages 427–431,Valencia, Spain, April 3-7, 2017*, Volume 2.

Thomas R Knapp. 1978. Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, 85(2):410.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations*.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.

Roderick P McDonald. 1970. Three common factor models for groups of variables. *Psychometrika*, 35(1):111–128.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Hannah Neuser. 2017. *Source Language of Lexical Transfer in Multilingual Learners: A Mixed Methods Approach*. Ph.D. thesis, Department of English, Stockholm University.

Aneta Pavlenko. 2009. Conceptual representation in the bilingual lexicon and second language vocabulary learning. *The bilingual mental lexicon: Interdisciplinary approaches*, pages 125–160.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. 2008. The matrix cookbook. *Technical University of Denmark*, 7(15):510.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *Proceedings of the 5th International Conference on Learning Representations*.

Charles Spearman. 1904. " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292.

Louis L Thurstone. 1931. Multiple factor analysis. *Psychological Review*, 38(5):406.

Michael E Tipping and Christopher M Bishop. 1999. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Ledyard R Tucker. 1958. An inter-battery method of factor analysis. *Psychometrika*, 23(2):111–136.

Uriel Weinreich. 1953. Languages in contact. findings andproblems. *New York: Linguistic Circle of New York and The Hague: Mouton*.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. 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*, pages 1006–1011.

A Joint Distribution

We show the form of the joint distribution for 2 views. Concatenating our data and parameters as below, we can use Equation (3) of (Ghahramani et al., 1996) to write

$$\boldsymbol{m} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}, \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_x \\ \boldsymbol{W}_y \end{bmatrix}$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi_x} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Psi_y} \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu_x} \\ \boldsymbol{\mu_y} \end{bmatrix}$$

$$p(\boldsymbol{m}, \boldsymbol{z}|\boldsymbol{\theta}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m} \\ \boldsymbol{z} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{0} \end{bmatrix}, \boldsymbol{\Sigma}_{m,z} \right) \quad (5)$$

$$\boldsymbol{\Sigma}_{m,z} = \begin{bmatrix} \boldsymbol{W}\boldsymbol{W}^\top + \boldsymbol{\Psi} & \boldsymbol{W} \\ \boldsymbol{W}^\top & \mathbb{I} \end{bmatrix}$$

It is clear that this generalises to any number of views of any dimension, as the concatenation operation does not make any assumptions.

B Projections to Latent Space $\mathbb{E}_{p(\boldsymbol{z}|\boldsymbol{x})}[\boldsymbol{z}]$

We can query the joint Gaussian in 5 using rules from (Petersen et al., 2008) Sections (8.1.2, 8.1.3) and we get

$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} \tilde{\boldsymbol{x}}, \mathbb{I} - \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} \boldsymbol{W}_x\right)$$
$$\mathbb{E}[\boldsymbol{z}|\boldsymbol{x}] = \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} \tilde{\boldsymbol{x}}$$

C Derivation for the Marginal Likelihood

We want to compute $p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$ so that we can then learn the parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_x, \boldsymbol{\theta}_y\}$, $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \boldsymbol{W}_i, \boldsymbol{\Psi}_i, \}$ by maximising the marginal likelihood as is done in Factor Analysis.

From the joint $p(\boldsymbol{m}, \boldsymbol{z}|\boldsymbol{\theta})$, again using rules from (Petersen et al., 2008) Sections (8.1.2) we get

$$\begin{aligned} p(\boldsymbol{m}|\boldsymbol{\theta}) &= p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) \\ &= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\Psi} \right) \end{aligned}$$

For the case of two views, the joint probability can be factored as

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) &= p(\boldsymbol{x}|\boldsymbol{\theta}_x) p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \\ p(\boldsymbol{x}|\boldsymbol{\theta}_x) &= \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \\ p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) &= \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{W}_y \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} \tilde{\boldsymbol{x}} + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}}\right) \\ &= \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{W}_y E[\boldsymbol{z}|\boldsymbol{x}] + \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}}\right), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_x &= \boldsymbol{W}_x \boldsymbol{W}_x^\top + \boldsymbol{\Psi}_x \\ \boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}} &= \boldsymbol{\Sigma}_y - \boldsymbol{W}_y \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} \boldsymbol{W}_x \boldsymbol{W}_y^\top \end{aligned}$$

D Scaled Reconstruction Errors

$$\begin{aligned} \log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}) &= \log p^*(\boldsymbol{x}|\boldsymbol{\theta}_x) + \log p^*(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \\ &\quad - \frac{1}{2}(\log |2\pi \boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{x}}| + \log |2\pi \boldsymbol{\Sigma}_{\boldsymbol{x}}|) \\ \log p^*(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) &= -\frac{1}{2} ||\tilde{\boldsymbol{y}} - \boldsymbol{W}_y E[\boldsymbol{z}|\boldsymbol{x}]||^2_{\boldsymbol{\Sigma}_{y|x}} \\ \log p^*(\boldsymbol{x}|\boldsymbol{\theta}_x) &= -\frac{1}{2} ||\boldsymbol{x} - \boldsymbol{\mu}_x||^2_{\boldsymbol{\Sigma}_x} \\ &= -\frac{1}{2} ||\boldsymbol{\Sigma}_x^{-\frac{1}{2}} \tilde{\boldsymbol{x}}||^2 \end{aligned}$$

Setting $\boldsymbol{A} = \boldsymbol{\Psi}_x \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Psi}_x$, we can re-parametrise as

$$\begin{aligned} \log p^*(\boldsymbol{x}|\boldsymbol{\theta}_x) &= -\frac{1}{2} ||\boldsymbol{\Psi}_x \boldsymbol{\Sigma}_x^{-1} \tilde{\boldsymbol{x}}||^2_{\boldsymbol{A}} \\ &= -\frac{1}{2} ||(\boldsymbol{\Sigma}_x - \boldsymbol{W}_x \boldsymbol{W}_x^\top) \boldsymbol{\Sigma}_x^{-1} \tilde{\boldsymbol{x}}||^2_{\boldsymbol{A}} \\ &= -\frac{1}{2} ||\tilde{\boldsymbol{x}} - \boldsymbol{W}_x \boldsymbol{W}_x^\top \boldsymbol{\Sigma}_x^{-1} \tilde{\boldsymbol{x}}||^2_{\boldsymbol{A}} \\ &= -\frac{1}{2} ||\tilde{\boldsymbol{x}} - \boldsymbol{W}_x E[\boldsymbol{z}|\boldsymbol{x}]||^2_{\boldsymbol{A}} \end{aligned}$$

E Expectation Maximisation for MBFA

Define

$$\tilde{\boldsymbol{x}} = \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{x}_v - \boldsymbol{\mu}_1 \end{bmatrix}, \boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 \\ \vdots \\ \boldsymbol{W}_v \end{bmatrix}$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}_1 & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{\Psi}_v \end{bmatrix} = \text{Bdiag}(\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_v)$$

Hence

$$p(\tilde{\boldsymbol{x}}|\boldsymbol{z}; \boldsymbol{\Psi}, \boldsymbol{W}) = \mathcal{N}(\tilde{\boldsymbol{x}}|\boldsymbol{W}\boldsymbol{z}, \boldsymbol{\Psi})$$

Method	EN-IT	IT-EN	EN-FR	FR-EN	IT-FR	FR-IT
MBFA-1K	71.9	73.3	76.7	78.2	82.4	77.5
MBFA-20K	71.9	73.4	76.7	78.1	82.6	77.5
MBFA-1K+CSLS	77.5	77.6	81.9	82.0	86.8	82.1
MBFA-20K+CSLS	77.4	77.7	81.9	82.1	86.8	81.9

Table 8: Precision @1 between MBFA fitted for 1K iterations and MBFA fitted for 20K iterations.

This follows the same form as regular factor analysis, but with a block-diagonal constraint on $\boldsymbol{\Psi}$. Thus by Equations (5) and (6) of (Ghahramani et al., 1996), we apply EM as follows.

E-Step: Compute $\mathbb{E}[\boldsymbol{z}|\boldsymbol{x}]$ and $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top|\boldsymbol{x}]$ given the parameters $\boldsymbol{\theta}_t = \{\boldsymbol{W}_t, \boldsymbol{\Psi}_t\}$.

$$\begin{aligned}\mathbb{E}[\boldsymbol{z}^{(i)}|\tilde{\boldsymbol{x}}^{(i)}] &= \boldsymbol{B}_t\tilde{\boldsymbol{x}}^{(i)} \\ \mathbb{E}[\boldsymbol{z}^{(i)}\boldsymbol{z}^{(i)\top}|\tilde{\boldsymbol{x}}^{(i)}] &= \mathbb{I} - \boldsymbol{B}_t\boldsymbol{W}_t + \boldsymbol{B}_t\tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top}\boldsymbol{B}_t^\top \\ &= \boldsymbol{M}_t + \boldsymbol{B}_t\tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top}\boldsymbol{B}_t^\top \qquad (6)\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{M}_t &= \left(\mathbb{I} + \boldsymbol{W}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{W}_t\right)^{-1} \\ \boldsymbol{B}_t &= \boldsymbol{W}_t^\top(\boldsymbol{\Psi}_t + \boldsymbol{W}_t\boldsymbol{W}_t^\top)^{-1} \\ &= \boldsymbol{M}_t\boldsymbol{W}_t^\top\boldsymbol{\Psi}_t^{-1}. \qquad (7)\end{aligned}$$

Equation 6 is obtained by applying the Woodbury identity, and Equation 7 by applying the closely related push-through identity, as found in Section 3.2 of (Petersen et al., 2008).

M-Step: Update parameters $\boldsymbol{\theta}_{t+1} = \{\boldsymbol{W}_{t+1}, \boldsymbol{\Psi}_{t+1}\}$.

Define

$$\boldsymbol{S} = \frac{1}{m}\sum_{i=1}^{m}\tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top}$$

By first observing

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\tilde{\boldsymbol{x}}^{(i)}\mathbb{E}[\boldsymbol{z}^{(i)}|\tilde{\boldsymbol{x}}^{(i)}]^\top &= \boldsymbol{S}\boldsymbol{B}_t^\top \\ \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}[\boldsymbol{z}^{(j)}\boldsymbol{z}^{(j)\top}|\tilde{\boldsymbol{x}}^{(j)}] &= \boldsymbol{M}_t + \boldsymbol{B}_t\boldsymbol{S}\boldsymbol{B}_t^\top,\end{aligned}$$

update the parameters as follows.

$$\begin{aligned}\boldsymbol{W}_{t+1} &= \boldsymbol{S}\boldsymbol{B}_t^\top\left(\mathbb{I} - \boldsymbol{B}_t\boldsymbol{W}_t + \boldsymbol{B}_t\boldsymbol{S}\boldsymbol{B}_t^\top\right)^{-1} \\ &= \boldsymbol{S}\boldsymbol{B}_t^\top\left(\boldsymbol{M}_t + \boldsymbol{B}_t\boldsymbol{S}\boldsymbol{B}_t^\top\right)^{-1} \\ \widetilde{\boldsymbol{\Psi}}_{t+1} &= \frac{1}{m}\sum_{i=1}^{m}\tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top} - \boldsymbol{W}_{t+1}\mathbb{E}[\boldsymbol{z}^{(i)}|\tilde{\boldsymbol{x}}^{(i)}]\tilde{\boldsymbol{x}}^{(i)\top} \\ &= \boldsymbol{S} - \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{W}_{t+1}\boldsymbol{B}_t\tilde{\boldsymbol{x}}^{(i)}\tilde{\boldsymbol{x}}^{(i)\top} \\ &= \boldsymbol{S} - \boldsymbol{W}_{t+1}\boldsymbol{B}_t\boldsymbol{S} \\ &= \boldsymbol{S} - \boldsymbol{S}\boldsymbol{B}_t^\top\boldsymbol{W}_{t+1}^\top\end{aligned}$$

Imposing the block diagonal constraint,

$$\boldsymbol{\Psi}_{t+1} = \text{Bdiag}\left((\widetilde{\boldsymbol{\Psi}}_{t+1})_{11}, \ldots, (\widetilde{\boldsymbol{\Psi}}_{t+1})_{vv}\right)$$

where $(\tilde{\boldsymbol{\Psi}})_{ii} = \boldsymbol{\Psi}_i$.

F Independence to Noise in Direct Methods

We are maximising the following quantity with respect to $\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{\Psi}\}$

$$\begin{aligned}p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}) &= \prod_i p(\boldsymbol{y}^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\theta}) \\ &= \prod_i \mathcal{N}(\boldsymbol{y}^{(i)}; \boldsymbol{W}\boldsymbol{x}^{(i)} + \boldsymbol{\mu}, \boldsymbol{\Psi}) \\ \log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}) &= -\frac{1}{2}\left(\sum_i \|\boldsymbol{y}^{(i)} - \boldsymbol{W}\boldsymbol{x}^{(i)}\|_{\boldsymbol{\Psi}}^2 - C\right)\end{aligned}$$

Then the partial derivative $\mathcal{Q} = \frac{\partial \log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta})}{\partial \boldsymbol{W}}$ is proportional to

$$\begin{aligned}\mathcal{Q} &\propto \left(\sum_i \boldsymbol{\Psi}^{-1}(\boldsymbol{y}^{(i)} - \boldsymbol{W}\boldsymbol{x}^{(i)})\boldsymbol{x}^{(i)\top}\right) \\ &\propto \boldsymbol{\Psi}^{-1}\left(\sum_i \boldsymbol{y}^{(i)}\boldsymbol{x}^{(i)\top} - \boldsymbol{W}\sum_i \boldsymbol{x}^{(i)}\boldsymbol{x}^{(i)\top}\right)\end{aligned}$$

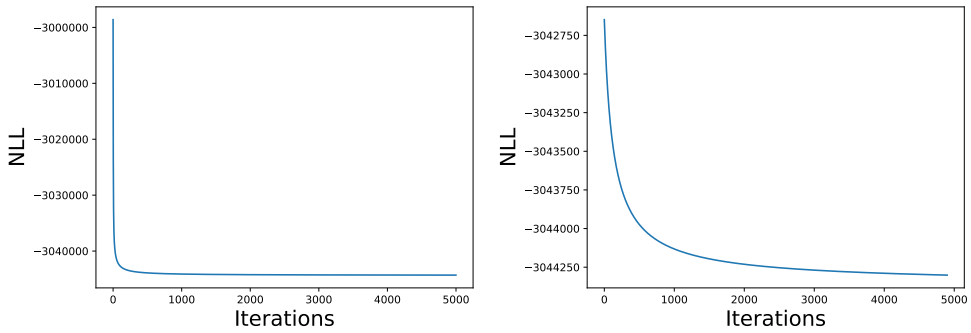

Figure 4: Training curve of EM algorithm over the first 5,000 iterations. It is clear that the procedure quickly finds a good approximation to the optimal parameters and then slowly converges to the real optimum. Left picture shows the entire training curve, while the right picture starts from iteration 100.

The maximum likelihood is achieved when

$$\frac{\partial \log p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta})}{\partial \boldsymbol{W}} = \mathbf{0},$$

and since $\boldsymbol{\Psi}^{-1}$ has an inverse (namely $\boldsymbol{\Psi}$), this means that

$$\boldsymbol{W} \sum_i \boldsymbol{x}^{(i)} \boldsymbol{x}^{(i)\top} = \sum_i \boldsymbol{y}^{(i)} \boldsymbol{x}^{(i)\top}$$

It is clear from here that the MLE of $\boldsymbol{W}$ does not depend on $\boldsymbol{\Psi}$, thus we can conclude that adding a noise parameter to this directed linear model has no effect on its predictions.

G Learning curve of EM

Figure 4 shows the negative log-likelihood of the three language model over the first 5,000 iterations. The precision of the learned model is very close when evaluated at iteration 1,000 and at iteration 20,000 as seen in Table 8. This suggests that the model need not be trained to full convergence to work well.