Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Survey Track

A Survey on Word Meta-Embedding Learning

Danushka Bollegala[1,2] , **James O'Neill**[1]
[1]University of Liverpool
[2]Amazon
{danushka, james.o-neill}@liverpool.ac.uk

Abstract

Meta-embedding (ME) learning is an emerging approach that attempts to learn more accurate word embeddings given existing (source) word embeddings as the sole input. Due to their ability to incorporate semantics from multiple source embeddings in a compact manner with superior performance, ME learning has gained popularity among practitioners in NLP. To the best of our knowledge, there exist no prior systematic survey on ME learning and this paper attempts to fill this need. We classify ME learning methods according to multiple factors such as whether they (a) operate on static or contextualised embeddings, (b) trained in an unsupervised manner or (c) fine-tuned for a particular task/domain. Moreover, we discuss the limitations of existing ME learning methods and highlight potential future research directions.

1 Introduction

Given independently trained multiple word representations (aka *embeddings*) learnt using diverse algorithms and lexical resources, word meta-embedding (ME) learning methods [Yin and Schütze, 2016; Bao and Bollegala, 2018; Bollegala *et al.*, 2018a; Wu *et al.*, 2020; He *et al.*, 2020; Jawanpuria *et al.*, 2020; Coates and Bollegala, 2018] attempt to learn more accurate and wide-coverage word embeddings. The input and output word embeddings to the ME algorithm are referred respectively as the *source* and *meta*-embeddings. ME has emerged as a promising ensemble approach to combine diverse pretrained source word embeddings to preserve their complementary strengths.

The problem settings of ME learning differ from that of source embedding learning in important ways as follows.

1. ME methods must be agnostic to the methods used to train the source embeddings.

The source embeddings used as the inputs to an ME learning method can be trained using different methods. For example, context-insensitive *static* word embedding methods [Dhillon *et al.*, 2015; Mnih and Hinton, 2009; Collobert *et al.*, 2011; Huang *et al.*, 2012; Mikolov *et al.*, 2013; Pennington *et al.*, 2014] represent a word by a single vector that does not vary depending on the context in which the word occurs. On the other hand, *contextualised* word embedding methods [Peters *et al.*, 2018; Devlin *et al.*, 2019; Yang *et al.*, 2019; Lan *et al.*, 2020; Liu *et al.*, 2019] represent the same word with different embeddings in its different contexts. It is not clear beforehand which word embedding is best for a particular NLP task. By being agnostic to the underlying differences in the source embedding learning methods, ME learning methods are in principle able to incorporate a wide range of source word embeddings. Moreover, this decoupling of source embedding learning from the ME learning simplifies the latter.

2. ME methods must not assume access to the original training resources used to train the source embeddings.

Source embeddings can be trained using different linguistic resources such as text corpora or dictionaries [Tissier *et al.*, 2017; Alsuhaibani *et al.*, 2019; Bollegala *et al.*, 2016]. Although pretrained word embeddings are often publicly released and are free of charge to use, the resources on which those embeddings were originally trained might not be publicly available due to copyright and licensing restrictions. Consequently, ME methods have not assumed the access to the original training resources that were used to train the source embeddings. Therefore, an ME method must obtain all semantic information of words directly from the source embeddings.

3. ME methods must be able to handle pretrained word embeddings of different dimensionalities.

Because ME methods operate directly on pretrained source word embeddings without retraining them on linguistic resources, the dimensionalities of the source word embeddings are often different. Prior work [Yin and Shen, 2018; Levy *et al.*, 2015] studying word embeddings have shown that the performance of a static word embedding is directly influenced by its dimensionality. ME learning methods use different techniques such as concatenation [Yin and Schütze, 2016], orthogonal projections [Jawanpuria *et al.*, 2020; He *et al.*, 2020] and averaging [Coates and Bollegala, 2018] after applying zero-padding to the sources with smaller dimensionalities as necessary to handle source embeddings with different dimensionalities.

Applications of ME: ME learning is attractive from an NLP practitioners point for several reasons. First, as mentioned above, there is already a large number of pretrained and publicly available repositories of static and contextialised word

embeddings. However, it is not readily obvious what is the best word embedding method to represent the input in a particular NLP application. We might not be able to try each and every source embedding due to time or computational constraints. ME learning provides a convenient alternative to selecting the single best word embedding, where we can use a ME trained from *all* available source embeddings. Second, *unsupervised ME learning methods* (§ 3) do *not* require labelled data when creating an ME from a given set of source word embeddings. This is particularly attractive in scenarios where we do not have sufficiently large training resources for learning word embeddings from scratch but have access to multiple pretrained word embeddings. Moreover, by using multiple source embeddings we might be able to overcome the limited vocabulary coverage in the individual sources. Third, in situations where there is some labelled data for the target task or domain, we can use *supervised ME learning methods* (§ 4) to fine-tune the MEs for the target task or domain.

From a theoretical point-of-view, ME learning can be seen as an instance of ensemble learning [Dietterich, 2002; Polikar, 2012], where we incorporate information from multiple models of lexical (word-level) semantics to learn an improved representation model. An ensemble typically helps to cancel out noise in individual models, while reinforcing the useful patterns repeated in multiple models [Muromägi *et al.*, 2017]. Although there are some theoretical work studying word embedding learning [Arora *et al.*, 2016; Mu *et al.*, 2018; Bollegala *et al.*, 2018b], the theoretical analysis of ME learning has been under-developed, with the exception of concatenated meta-embeddings [Bollegala, 2022]. For example, under what conditions can we learn a better ME than individual source embeddings is an important theoretical consideration. ME learning can also been seen as related to *model distillation* [Passos *et al.*, 2018; Hinton *et al.*, 2015] where we must learn a simpler *student* model from a more complicated *teacher* model. Model distillation is an actively researched topic in deep learning where it is attractive to learn smaller networks involving a lesser number of parameters from a larger network to avoid overfitting and inference-time efficiency.

In this survey paper we focus on word-level ME learning. We first define the ME problem (§2) and cover unsupervised (§ 3) and supervised (§ 4) ME methods. We also look at multilingual ME in §5. Finally, we discuss the performance of different ME methods (§6.1) and present potential future research directions (§6.2). Moreover, we publicly release a ME framework[1] that implements several ME methods covered in this paper, which we believe would be useful to further enhance the readers' understanding on this emerging topic.

2 Meta-Embedding – Problem Definition

Let us consider a set of N source word embeddings $s_1, s_2, \ldots, s_N$ respectively covering vocabularies (i.e. sets of words) $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_n$. The embedding of a word w in s_j is denoted by $\boldsymbol{s}_j(w) \in \mathbb{R}^{d_j}$, where d_j is the dimensionality of s_j. We can represent s_j by an embedding matrix $\mathbf{E}_j \in \mathbb{R}^{d_j \times |\mathcal{V}_j|}$. For example, $\mathbf{E}_1$ could be the embedding matrix obtained by running skip-gram with negative sampling (SGNS) [Mikolov *et al.*, 2013] on a corpus, whereas $\mathbf{E}_2$ could be that obtained from global vector prediction (GloVe) [Pennington *et al.*, 2014] etc. Then, the problem of ME can be stated as – *what is the optimal way to combine* $\mathbf{E}_1, \ldots, \mathbf{E}_n$ *such that some goodness measure defined for the accuracy of the semantic representation for the words is maximised?*

The source word embeddings in general do not have to cover the same set of words. If $w \notin \mathcal{V}_n$, we can either assign a zero embedding or a random embedding as $\boldsymbol{s}_n(w)$ as a workaround. [Yin and Schütze, 2016] proposed a method to predict source embeddings for the words missing in a particular source. Specifically, given two different sources s_n and s_m (where $n \neq m$) they learn a projection matrix $\mathbf{A} \in \mathbb{R}^{d_m \times d_n}$ using the words in $\mathcal{V}_n \cap \mathcal{V}_m$, the intersection between the vocabularies covered by both s_n and s_m. They find $\mathbf{A}$ by minimising the sum of squared loss, $\sum_{w \in \mathcal{V}_n \cap \mathcal{V}_m} ||\mathbf{A}\boldsymbol{s}_n(w) - \boldsymbol{s}_m(w)||_2^2$. Finally, we can predict the source embedding for a word $w' \notin \mathcal{V}_m$ and $w' \in \mathcal{V}_n$ using the learnt $\mathbf{A}$ as $\mathbf{A}\boldsymbol{s}_n(w')$. If we have multiple sources, we can learn such projection matrices between each pair of sources and in both directions. We can then, for example, consider the average of all predicted embeddings for a word as its source embedding in a particular source. After this preprocessing step, all words will be covered by all source embeddings. [Yin and Schütze, 2016] showed that by applying this preprocessing step prior to learning MEs (referred to as the 1TON+ method in their paper) to significantly improve the performance of the learnt MEs. However, as we see later, much prior work in ME learning do assume a common vocabulary over all source embeddings for simplicity. Without loss of generality, we will assume that all words are covered by a common vocabulary $\mathcal{V}$ after applying any one of the above-mentioned methods.

3 Unsupervised Meta-Embedding Learning

In unsupervised ME we do not assume the availability of any manually-annotated labelled data that we can use in the learning process. In this setting all data that we have at our disposal is limited to the pretrained source embeddings.

3.1 Concatenation

One of the simplest approaches to create an ME under the unsupervised setting is vector concatenation [Bao and Bollegala, 2018; Yin and Schütze, 2016; Bollegala *et al.*, 2018a]. Denoting concatenation by $\oplus$, we can express the concatenated ME, $\boldsymbol{m}_{\text{conc}}(w) \in \mathbb{R}^{d_1 + \ldots + d_N}$, of a word $w \in \mathcal{V}$ by (1).

$$\begin{aligned} \boldsymbol{m}_{\text{conc}}(w) &= \boldsymbol{s}_1(w) \oplus \ldots \boldsymbol{s}_N(w) \\ &= \oplus_{j=1}^{N} \boldsymbol{s}_j(w) \end{aligned} \tag{1}$$

Goikoetxea *et al.* [2016] showed that the concatenation of word embeddings learnt separately from a corpus and WordNet to produce superior word embeddings. However, one disadvantage of using concatenation to produce MEs is that it increases the dimensionality of the ME space, which is the sum of the dimensionalities of the sources. [Yin and Schütze, 2016] post-processed the MEs created by concatenating the source embeddings using Singular Value Decomposition (SVD) to

[1] https://github.com/Bollegala/Meta-Embedding-Framework

reduce the dimensionality. However, applying SVD often results in degradation of accuracy in the MEs compared to the original concatenated version [Bollegala *et al.*, 2018a].

It is easier to see that concatenation does not remove any information that is already covered by the source embeddings. However, it is not obvious under what conditions concatenation could produce an ME that is superior to the input source embeddings. Bollegala [2022] shows that concatenation minimises the pairwise inner-product (PIP) [Yin and Shen, 2018] loss between the source embeddings and an idealised ME. PIP loss has been shown to be directly related to the dimensionality of a word embedding, and has been used as a criterion for selecting the optimal dimensionality for static word embeddings. They propose a weighted variant of concatenation where the dimensions of each source is linearly weighted prior to concatenation. The weight parameters can be learnt in an unsupervised manner by minimising the empirical PIP loss.

3.2 Averaging

Note that as we already mentioned in §1, source embeddings are trained independently and can have different dimensionalities. Even when the dimensionalities do agree, vectors that lie in different vector spaces cannot be readily averaged. However, rather surprisingly, [Coates and Bollegala, 2018] showed that accurate MEs can be produced by first zero-padding source embeddings as necessary to bring them to a common dimensionality and then by averaging to create $\boldsymbol{m}_{\text{avg}}(w)$ as given by (2).

$$\boldsymbol{m}_{\text{avg}}(w) = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{s}_j^*(w) \tag{2}$$

Here, $\boldsymbol{s}_j^*(w)$ is the zero-padded version of $\boldsymbol{s}_j(w)$ such that its dimensionality is equal to $\max(d_1, \ldots, d_N)$. In contrast to concatenation, averaging has the desirable property that the dimensionality of the ME is upper-bounded by $\max(d_1, \ldots, d_N) < \sum_{j=1}^{N} d_j$. Coates and Bollegala [2018] showed that when word embeddings in each source are approximately orthogonal, a condition that they empirically validate for pre-trained word embeddings, averaging can approximate the MEs created by concatenating.

Although averaging does not increase the dimensionality of the ME space as with concatenation, it does not consistently outperform concatenation, especially when the orthogonality condition does not hold. To overcome this problem, Jawanpuria *et al.* [2020] proposed to first learn orthogonal projection matrices for each source embedding space. They measure the Mahalanobis metric between the projected source embeddings, which is a generalisation of the inner-product that does not assume the dimensions in the vector space to be uncorrelated.

To explain their proposal further, let us consider two sources s_1 and s_2 with identical dimensionality d. Let us assume the orthogonal projection matrices for s_1 and s_2 to be respectively $\mathbf{A}_1 \in \mathbb{R}^{d\times d}$ and $\mathbf{A}_2 \in \mathbb{R}^{d\times d}$. The two words $w_i, w_j \in \mathcal{V}_1 \cap \mathcal{V}_2$ are projected to a common space respectively as $\mathbf{A}_1\boldsymbol{s}_1(w_i) \in \mathbb{R}^d$ and $\mathbf{A}_2\boldsymbol{s}_2(w_j) \in \mathbb{R}^d$. The similarity in this projected space is computed using a Mahalanobis metric $(\mathbf{A}_1\boldsymbol{s}_1(w_i))^\top \mathbf{B}(\mathbf{A}_2\boldsymbol{s}_2(w_j))$ defined by the matrix $\mathbf{B} \in \mathbb{R}^{d\times d}$. They learn $\mathbf{A}_1, \mathbf{A}_2$ and $\mathbf{B}$ such that the above metric computed between the projected source embeddings of the same word is close 1, while that for two different words is close to 0. Their training objective can be written concisely as in (3) using the embedding matrices $\mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{d\times|\mathcal{V}_1\cap\mathcal{V}_2|}$ and a matrix $\mathbf{Y}$ where the (i,j) element $Y_{ij} = 1$ if $w_i = w_j$ and 0 otherwise.

$$\underset{\mathbf{A}_1,\mathbf{A}_2,\mathbf{B}}{\text{minimise}} \left|\left|{\mathbf{E}_1}^\top {\mathbf{A}_1}^\top \mathbf{B}\mathbf{A}_2\mathbf{E}_2 - \mathbf{Y}\right|\right|_F^2 + \lambda\,||\mathbf{B}||_F^2 \tag{3}$$

Here, $\lambda \geq 0$ is a regularisation coefficient corresponding to the Frobenius norm regularisation of $\mathbf{B}$, which prefers smaller Mahalanobis matrices. They show that the averaging of the projected source embeddings (i.e. $(\mathbf{B}^{\frac{1}{2}}\mathbf{A}_1\boldsymbol{s}_1(w) + \mathbf{B}^{\frac{1}{2}}\mathbf{A}_2\boldsymbol{s}_2(w))/2$) to outperform simple non-projected averaging (given by (2)). Learning such orthogonal projections for the sources has shown to be useful even in supervised ME learning [He *et al.*, 2020] as discussed later in §4.

3.3 Linear Projections

In their pioneering work on ME, Yin and Schütze [2016] proposed to project source embeddings to a common space via source-specific linear transformations, which they refer to as 1TON. They require that the ME of a word w, $\boldsymbol{m}_{1\text{TON}}(w) \in \mathbb{R}^{d_m}$, reconstruct each source embedding, $\boldsymbol{s}_j(w)$ of w using a linear projection matrix, $\mathbf{A}_j \in \mathbb{R}^{d_j\times d_m}$, from s_j to the ME space by as given by (4).

$$\hat{\boldsymbol{s}}_j(w) = \mathbf{A}_j\boldsymbol{m}_{1\text{TON}}(w) \tag{4}$$

Here, $\hat{\boldsymbol{s}}_j(w)$ is the reconstructed source embedding of w from the ME. Next, the squared Euclidean distance between the source- and MEs is minimised over all words in the intersection of the source vocabularies, subjected to Frobenius norm regularisation as in (5).

$$\underset{\substack{\forall_{j=1}^N \mathbf{A}_j \\ \forall_{w\in\mathcal{V}}\ \boldsymbol{m}_{1\text{TON}}(w)}}{\text{minimise}} \sum_{j=1}^{N} \alpha_j \left(\sum_{w\in\mathcal{V}} ||\hat{\boldsymbol{s}}_j(w) - \boldsymbol{s}_j(w)||_2^2 + ||\mathbf{A}_j||_F^2\right) \tag{5}$$

They use different weighting coefficients α_j to account for the differences in accuracies of the sources. They determine α_j using the Pearson correlation coefficients computed between the human similarity ratings and cosine similarity computed using the each source embedding between word pairs on the [Miller and Charles, 1998] dataset. The parameters can be learnt using stochastic gradient descent, alternating between projection matrices and MEs.

Muromägi *et al.* [2017] showed that by requiring the projection matrices to be orthogonal (corresponding to the Orthogonal Procrustes Problem) the accuracy of the learnt MEs is further improved. However, 1TON requires all words to be represented in all sources. To overcome this limitation they predict the source embedding for missing words as described in §2.

Assuming that a single *global* linear projection can be learnt between the ME space and each source embedding as done by Yin and Schütze [2016] is a stronger requirement. Bollegala *et al.* [2018a] relaxed this requirement by learning *locally linear* (LLE) MEs. To explain this method further let us consider computing the LLE-based ME, $\boldsymbol{m}_{\text{LLE}}(w)$, of a word

$w \in \mathcal{V}_1 \cap \mathcal{V}_2$ using two sources s_1 and s_2. First, they compute the set of nearest neighbours, $\mathcal{N}_j(w)$, of w in s_j and represent w as the linearly-weighted combination of its neighbours by a matrix $\mathbf{A}$ by minimisng (6).

$$\underset{\mathbf{A}}{\text{minimise}} \sum_{j=1}^{2} \sum_{w \in \mathcal{V}_1 \cap \mathcal{V}_2} \left|\left| \boldsymbol{s}_j(w) - \sum_{w' \in \mathcal{N}_j(w)} A_{ww'} \boldsymbol{s}_j(w) \right|\right|_2^2 \quad (6)$$

They use AdaGrad to find the optimal $\mathbf{A}$. Next, MEs are learnt by minimising (7) using the learnt neighbourhood reconstruction weights in $\mathbf{A}$ are preserved in a vector space common to all source embeddings.

$$\sum_{w \in \mathcal{V}_1 \cap \mathcal{V}_2} \left|\left| \boldsymbol{m}_{\text{LLE}}(w) - \sum_{j=1}^{2} \sum_{w' \in \mathcal{N}_j(w)} C_{ww'} \boldsymbol{m}_{\text{LLE}}(w') \right|\right|_2^2 \quad (7)$$

Here, $C_{ww'} = A_{ww'} \sum_{j=1}^{2} \mathbb{I}[w' \in \mathcal{N}_j(w)]$, where $\mathbb{I}$ is the indicator function which returns 1 if the statement evaluated is True. Optimal MEs can then be found by solving an eigen-decomposition of the matrix $(\mathbf{I} - \mathbf{C})^\top (\mathbf{I} - \mathbf{C})$, where $\mathbf{C}$ is the matrix formed by arranging $C_{ww'}$ as the (w, w') element. This approach has the advantage that it does not require all words to be represented by all sources, thereby obviating the need to predict missing source embeddings prior to ME.

3.4 Autoencoding

Bao and Bollegala [2018] modelled ME learning as an *autoencoding* problem where information embedded in different sources are integrated at different levels to propose three types of MEs: Decoupled Autoencoded ME (DAEME) (independently encode each source and concatenate), Concatenated Autoencoded ME (CAEME) (independently decode MEs to reconstruct each source), and Averaged Autoencoded ME (AAEME) (similar to DAEME but instead of concatenation uses averaging). Given the space constraints we describe only the AAEME model, which was the overall best performing among the three.

Consider two sources s_1 and s_2, which are encoded respectively by two encoders E_1 and E_2. The AAEME of w is computed as the ℓ_2 normalised average of the encoded source embeddings as given by (8).

$$\boldsymbol{m}_{\text{AAEME}}(w) = \frac{E_1(\boldsymbol{s}_1(w)) + E_2(\boldsymbol{s}_2(w))}{||E_1(\boldsymbol{s}_1(w)) + E_2(\boldsymbol{s}_2(w))||_2} \quad (8)$$

Two independent decoders, D_1 and D_2, are trained to reconstruct the two sources from the ME. E_1, E_2, D_1 and D_2 are jointly learnt to minimise the weighted reconstruction loss given by (9).

$$\underset{E_1,E_2,D_1,D_2}{\text{minimise}} \sum_{w \in \mathcal{V}_1 \cap \mathcal{V}_2} (\lambda_1 ||\boldsymbol{s}_1(w) - D_1(E_1(\boldsymbol{s}_1(w)))||_2^2 + \lambda_2 ||\boldsymbol{s}_2(w) - D_2(E_2(\boldsymbol{s}_2(w)))||_2^2) \quad (9)$$

The weighting coefficients λ_1 and λ_2 can be used to assign different emphasis to reconstructing the two sources and are tuned using a validation dataset. In comparison to methods that learn globally or locally linear transformations [Bollegala *et al.*, 2018a; Yin and Schütze, 2016], autoencoders learn non-linear transformations. Their proposed autoencoder variants outperform 1TON and 1TON+ on multiple benchmark tasks.

Although our focus in this survey is word-level ME learning, sentence-level ME methods have also been proposed [Poerner *et al.*, 2020; Takahashi and Bollegala, 2022]. Poerner *et al.* [2020] proposed several methods to combine sentence-embeddings from pretrained encoders such as by concatenating and averaging the individual sentence embeddings.These methods correspond to using sentence embeddings instead of source word embeddings in (1) and (2) with ℓ_2 normalised sources. Moreover, they used the Generalised Canonical Correlation Analysis (GCCA), which extends Canonical Correlation Analysis to more than three random vectors, to learn sentence-level MEs. They also extend AAEME method described in §3.4 to multiple sentence encoders, where they learn an autoencoder between each pair of sources. They found that GCCA to perform best in sentence similarity prediction tasks. Takahashi and Bollegala [2022] proposed an unsupervised sentence-level ME method, which learns attention weights and transformation matrices over contextualised embeddings such that multiple word- and sentence-level co-occurrence criteria are simultaneously satisfied.

4 Supervised Meta-Embedding Learning

MEs have also been learned specifically for a set of supervised tasks. Unlike unsupervised ME learning methods described in §3, supervised MEs use end-to-end learning and fine-tune the MEs specifically for the downstream tasks of interest.

O'Neill and Bollegala [2018] used MEs to regularise a supervised learner by reconstructing the ensemble set of pretrained embeddings as an auxiliary task to the main supervised task whereby the encoder is shared between both tasks. (10) shows the auxiliary reconstruction mean squared error loss weighted by α for each word w_t in a sequence of of length T words, and the main sequence classification task cross-entropy loss, weighted by β. Here, $f_{\theta_{\text{aux}}}$ is the subnetwork of f_θ that corresponds to the ME reconstruction and $f_{\theta_{\text{main}}}$ is the subnetwork used to learn on the main task, i.e f_θ except the decoder layer of the ME autoencoder.

$$\underset{\theta}{\text{minimise}} \frac{1}{TN} \sum_{t=1}^{T} \Big[\alpha\big(\mathbf{f}_{\theta_{\text{aux}}}(\boldsymbol{m}_{\text{conc}}(w_t)) - \boldsymbol{m}_{\text{conc}}(w_t)\big)^2 + \beta \sum_{c=1}^{C} \mathbf{f}_{\theta_{\text{main}}}\big(\boldsymbol{m}_{\text{conc}}(w_t)\big) \log y_{t,c}\Big] \quad (10)$$

The ME reconstruction shows improved performance on both intrinsic tasks (word similarity and relatedness) and extrinsic tasks (named entity recognition, part of speech tagging and sentiment analysis). They also show that MEs require less labelled data for Universal Part of Speech tagging[2] to perform as well as unsupervised MEs. Wu *et al.* [2020] also successfully deploy the aforementioned ME regularisation in the supervised learning setting. They showed that as the ME hidden

[2]https://universaldependencies.org/u/pos/all.html

layer becomes smaller, the improvement in performance for supervised MEs over unsupervised MEs becomes larger.

Kiela *et al.* [2018] proposed a supervised ME learning method where they compute the ME of a word using a dynamically-weighted sum of the projected source word embeddings. Each source embedding, s_j, is projected to a common d-dimensional space using a matrix $\mathbf{P}_j \in \mathbb{R}^{d_j \times d}$ and a basis vector $\boldsymbol{b}_j \in \mathbb{R}^d$ as given by (11).

$$\boldsymbol{s}'_j(w) = \mathbf{P}_j \boldsymbol{s}_j(w) + \boldsymbol{b}_j \tag{11}$$

Next, the Dynamic Meta Embedding, $\boldsymbol{m}_{\text{dme}}(w)$, of a word w is computed as in (12).

$$\boldsymbol{m}_{\text{dme}}(w) = \sum_{j=1}^{N} \alpha_{w,j} \boldsymbol{s}'_j(w) \tag{12}$$

Here, $\alpha_{w,j}$ is the scalar weight associated with w in source s_j, and is computed via self-attention mechanism as given by (13).

$$\alpha_{w,j} = \phi(\boldsymbol{a}^\top \boldsymbol{s}'_j(w) + b) \tag{13}$$

Here, ϕ is an activation function such as softmax, $\boldsymbol{a} \in \mathbb{R}^d$ and $b \in R$ are respectively source and word independent attention parameters to be learnt using labelled data. They also proposed a contextualised version of DME (CDME), where they used the hidden state from a BiLSM that takes the projected source embeddings as the input. Although the differences between DME and CDME were not statistically significant, overall, the highest maximum performances were reported with CDME.

Xie *et al.* [2019] extended this approach by introducing task-specific factors for a downstream task by computing the pairwise interactions between embeddings in the ensemble set. He *et al.* [2020] also create task-specific MEs by learning orthogonal transformations to linearly combine the source embeddings. As already discussed in §3.2, enforcing orthogonality on the transformation matrix has shown to improve performance of ME.

Lange *et al.* [2021] proposed feature-based adversarial meta-embeddings (FAME) for creating MEs from both word- and subword-level source embeddings. Specifically, a given token is represented by source embeddings as well as source-independent features related to the token such as frequency, length, shape (e.g. upper/lowercasing, punctuation, number etc.). Source embeddings are projected to a common vector space using linear projections, and their attention-weighted sum is computed as the ME. A gradient reversal layer [Ganin *et al.*, 2016] is used to learn the projection and attention related parameters. FAME achieves SoTA results for part-of-speech (POS) tagging in 27 languages.

5 Multi-lingual Meta-Embedding Learning

MEs have also been extended to the cross-lingual and multi-lingual settings. Winata *et al.* [2019a] use self-attention across different embeddings to learn multi-lingual MEs. For Named Entity Recognition (NER), the multi-lingual embeddings in the ensemble set are concatenated and passed as the input into a transformer encoder and a conditional random field is used as the classification layer. The use of self-attention weights showed to improve over a linear layer without self-attention weights. Winata *et al.* [2019b] have also extended this to using hierarchical MEs (HMEs), which refers to word-level MEs that are created via a weighted average of sub-word level ME representations. They find that HMEs outperform regular MEs on code-switching NER through the use of pretrained subword embeddings given by `fasttext` [Joulin *et al.*, 2017].

García *et al.* [2020] learn MEs for cross-lingual tasks by projecting the embeddings into a common semantic space. Embeddings of resource-rich languages can then be used to improve the quality of learned embeddings of low-resourced languages. This is achieved in three steps: (1) align the vector spaces of different vocabularies of each language using the bilingual mapper `VecMap` [Artetxe *et al.*, 2016], (2) create new embeddings for missing words in the source embeddings and (3) average the embeddings in the ensemble set. The resulting ME vocabulary will then be the union of the vocabularies of the word embeddings used. This method is referred to as MVM (Meta-VecMap).

Doval *et al.* [2018] align embeddings into a bilingual vector space using `VecMap` and `MUSE` [Conneau *et al.*, 2017] and use a linear layer to transform the aligned embeddings such that the average word vector in the ensemble set can be predicted in the target language from the source. This is motivated by the finding that the average of word embeddings is a good approximation for MEs [Coates and Bollegala, 2018] as already discussed in §3.2.

6 Discussion

In this section, we first discuss the performance of the different ME methods described in this paper. Next, we discuss the limitations in existing ME learning methods and highlight potential future research directions.

6.1 Evaluation and Performance

Given that ME are representing the meaning of words using vectors, MEs have been evaluated following the same benchmark tasks commonly used for evaluating source (static) word embeddings such as word or sentence similarity measurement, analogy detection, text classification, textual entailment recognition, part-of-speech tagging [Lange *et al.*, 2021], etc.

The performance of an ME created using a set of source embeddings depends on several factors such as the source embeddings used (e.g, the resources used to train the sources, their dimensionalities), dimensionality of the ME, and hyper-parameters (and how they were tuned) for the downstream tasks (in the case of supervised ME). However, prior work in ME learning have used different settings, which makes it difficult to make direct comparisons among results reported in different papers [García-Ferrero *et al.*, 2021].

Consistently across published results, concatenation has shown to be a competitive baseline, and averaging does not always outperform concatenation. Predicting source embeddings for out-of-vocabulary words has reported mixed results [García-Ferrero *et al.*, 2021]. Methods for predicting source embeddings for missing embeddings in a source uses simple linear transformations such as learning projection matrices [Yin and Schütze, 2016; García-Ferrero *et al.*, 2021].

However, whether such transformations always exist between independently trained vector spaces is unclear [Bollegala *et al.*, 2017]. On the other hand, averaging has reported good performance without requiring the prediction of missing source embeddings because average is already a good approximation for the missing source embeddings. However, scaling each dimension of a source embedding to zero mean and subsequently normalising the embeddings to unit ℓ_2 norm is required when the dimensionalities and norms of the source embeddings that are averaged are significantly different.

Moreover, carefully weighting sources using some validation data has shown to improve performance of concatenation [Yin and Schütze, 2016]. Although applying SVD reduces the dimensionality of the concatenated MEs, it does not always outperform the concatenation baseline. In particular, the number of singular values remains an important factor that influences the performance of this method [Bollegala, 2022]. The best performance for unsupervised ME has been reported by autoencoding methods, and in particular by AEME [Bao and Bollegala, 2018]. Overall, supervised or task-specific ME learning methods have reported superior performances over unsupervised ones [Lange *et al.*, 2021] in tasks such as sentence classification, POS tagging and NER. Therefore, when there is some labelled data available for the target task, we recommend using supervised ME methods. However, it remains unclear whether a supervised ME trained for one particular task will still perform well for a different task.

6.2 Issues and Future Directions

We outline issues and potential research directions in ME.

Contextualised MEs. Despite the good intrinsic performance, García-Ferrero *et al.* [2021] showed that none of the ME methods outperformed `fasttext` source embedding on GLUE benchmarks [Wang *et al.*, 2018]. Moreover, concatenation (with centering and normalising of source embeddings) and averaging have turned out to be strong baselines, often outperforming more complex ME methods. Given that contextualised embeddings obtain the SoTA performances on such extrinsic evaluations, we believe it is important for future research in ME learning to consider contextualised embeddings as sources instead of static word embeddings [Takahashi and Bollegala, 2022; Poerner *et al.*, 2020].

Sense-specific MEs. Thus far, ME learning methods have considered a word (or a sentence) as the unit of representation. However, the meaning of a word can be ambiguous and it can have multiple senses. Sense-embedding learning methods learn multi-prototype embeddings [Reisinger and Mooney, 2010; Neelakantan *et al.*, 2014] corresponding to the different senses of the same ambiguous word. How to combine sense embeddings with word embeddings to create sense-specific MEs remains an open research problem.

MEs For Sequence-to-Sequence Models. MEs have yet to be used for sequence-to-sequence generation tasks such as machine translation Therefore, we predict that a study of how they can be used in tandem with SoTA models such as the Transformer [Vaswani *et al.*, 2017] would be an impactful contribution. Particular questions of interests would be: (1) *How do Transformers perform on text generation tasks when preinitialised with MEs*? and (2) *Does a smaller model preinitialised with MEs outperform a model without MEs*?. Answering the aforementioned questions gives a clear indication of how MEs can be retrofitted into SoTA models and how they can obtain near SoTA results with shallower models.

Mitigating Negative Transfer. Creating an ME using all sources could lead to *negative transfer* [Pan and Yang, 2010] and consequently degrade model performance. Although attention-based source-weighting methods have been proposed [Xie *et al.*, 2019; Winata *et al.*, 2019b], that learn different weights for the sources, a systematic analysis of how negative transfer affects ME is required.

Theoretical Analysis of MEs: Compared to the empirical success, theoretical understanding of ME learning remains under-developed. Some of the important theoretical questions are: (1) *What is the optimal dimensionality for MEs to balance the memory vs performance tradeoff*?, (2) *What is the relative contribution between sources vs. ME learning algorithm towards the performance gains in downstream tasks*?, and (3) *What are the generalisation bounds for the performance of ME learning algorithms beyond a specific set of sources*?.

Social bias in MEs. Word embeddings have shown to encode worrying levels of unfair social biases such as gender, racial and religious biases [Kaneko and Bollegala, 2019; Bolukbasi *et al.*, 2016]. Given that MEs are incorporating multiple sources and further improve the accuracy of the embeddings, an unaddressed concern is whether an ME learning method would also amplify social biases contained in the source embeddings. It would be ethically inappropriate to deploy MEs with social biases to downstream NLP applications. We believe that further research is needed to detect and mitigate such biases in MEs.

7 Conclusion

We presented a survey of ME learning methods. We classified prior work into different categories such as unsupervised, supervised, sentence-level and multi-lingual ME learning methods. Finally, we highlighted potential future research directions. Given that ME learning is an active research topic, we hope this survey facilitates newcomers on this topic as well as providing inspiration to future developments in the broader AI community, incorporating existing word/text representations to create more accurate versions.

References

[Alsuhaibani *et al.*, 2019] Mohammed Alsuhaibani, Takanori Maehara, and Danushka Bollegala. Joint learning of hierarchical word embeddings from a corpus and a taxonomy. In *Proc. of AKBC*, 2019.

[Arora *et al.*, 2016] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *TACL*, 4:385–399, 2016.

[Artetxe *et al.*, 2016] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of

word embeddings while preserving monolingual invariance. In *Proc. of EMNLP*, pages 2289–2294, 2016.

[Bao and Bollegala, 2018] Cong Bao and Danushka Bollegala. Learning word meta-embeddings by autoencoding. In *Proc. of COLING*, pages 1650–1661, 2018.

[Bollegala *et al.*, 2016] Danushka Bollegala, Alsuhaibani Mohammed, Takanori Maehara, and Ken-ichi Kawarabayashi. Joint word representation learning using a corpus and a semantic lexicon. In *Proc. of AAAI*, pages 2690–2696, 2016.

[Bollegala *et al.*, 2017] Danushka Bollegala, Kohei Hayashi, and Ken-ichi Kawarabayashi. Learning linear transformations between counting-based and prediction-based word embeddings. *PLoS ONE*, 12(9):1–21, September 2017.

[Bollegala *et al.*, 2018a] Danushka Bollegala, Kohei Hayashi, and Ken-ichi Kawarabayashi. Think globally, embed locally — locally linear meta-embedding of words. In *Proc. of IJCAI-EACI*, pages 3970–3976, 2018.

[Bollegala *et al.*, 2018b] Danushka Bollegala, Yuichi Yoshida, and Ken-ichi Kawarabayashi. Using k-way Co-occurrences for Learning Word Embeddings. In *Proc. of AAAI*, pages 5037–5044, 2018.

[Bollegala, 2022] Danushka Bollegala. Learning meta word embeddings by unsupervised weighted concatenation of source embeddings. In *Proc. of IJCAI*, 2022.

[Bolukbasi *et al.*, 2016] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proc. of NeurIPS*, pages 4349–4357, 2016.

[Coates and Bollegala, 2018] Joshua Coates and Danushka Bollegala. Frustratingly easy meta-embedding – computing meta-embeddings by averaging source word embeddings. In *Proc. of NAACL-HLT*, pages 194–198, 2018.

[Collobert *et al.*, 2011] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuska. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493 – 2537, 2011.

[Conneau *et al.*, 2017] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, pages 4171–4186, 2019.

[Dhillon *et al.*, 2015] Paramveer S. Dhillon, Dean P. Foster, and Lyle H. Ungar. Eigenwords: Spectral word embeddings. *Journal of Machine Learning Research*, 16:3035–3078, 2015.

[Dietterich, 2002] Thomas Dietterich. *The Handbook of Brain Theory and Neural Networks*. MIT press Cambridge, 2002.

[Doval *et al.*, 2018] Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. Improving cross-lingual word embeddings by meeting in the middle. *arXiv preprint arXiv:1808.08780*, 2018.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *JMLR*, 17:1–35, 2016.

[García *et al.*, 2020] Iker García, Rodrigo Agerri, and German Rigau. A common semantic space for monolingual and cross-lingual meta-embeddings. *arXiv preprint arXiv:2001.06381*, 2020.

[García-Ferrero *et al.*, 2021] Iker García-Ferrero, Rodrigo Agerri, and German Rigau. Benchmarking meta-embeddings: What works and what does not. In *Findings of EMNLP*, pages 3957–3972, Punta Cana, Dominican Republic, 2021.

[Goikoetxea *et al.*, 2016] Josu Goikoetxea, Eneko Agirre, and Aitor Soroa. Single or multiple? combining word representations independently learned from text and wordnet. In *Proc. of AAAI*, pages 2608–2614, 2016.

[He *et al.*, 2020] Jingyi He, KC Tsiolis, Kian Kenyon-Dean, and Jackie Chi Kit Cheung. Learning Efficient Task-Specific Meta-Embeddings with Word Prisms. In *Proc. of COLING*, 2020.

[Hinton *et al.*, 2015] Geoffery E. Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv*, 10.48550/ARXIV.1503.02531, 2015.

[Huang *et al.*, 2012] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Proc. of ACL*, pages 873–882, 2012.

[Jawanpuria *et al.*, 2020] Pratik Jawanpuria, Satya Dev N T V, Anoop Kunchukuttan, and Bamdev Mishra. Learning geometric word meta-embeddings. In *Proc. of RepL4NLP*, pages 39–44, 2020.

[Joulin *et al.*, 2017] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of Tricks for Efficient Text Classification. In *Proc. of EACL*, pages 427–431, July 2017.

[Kaneko and Bollegala, 2019] Masahiro Kaneko and Danushka Bollegala. Gender-preserving debiasing for pre-trained word embeddings. In *Proc. of ACL*, pages 1641–1650, 2019.

[Kiela *et al.*, 2018] Douwe Kiela, Changhan Wang, and Kyunghyun Cho. Dynamic meta-embeddings for improved sentence representations. In *Proc. of EMNLP*, pages 1466–1477, 2018.

[Lan *et al.*, 2020] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *Proc. of ICLR*, 2020.

[Lange *et al.*, 2021] Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. FAME: Feature-based ad-

versarial meta-embeddings for robust input representations. In *Proc. of EMNLP*, pages 8382–8395, 2021.

[Levy *et al.*, 2015] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of Association for Computational Linguistics*, 3:211–225, 2015.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 10.48550/ARXIV.1907.11692, 2019.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, and Jeffrey Dean. Efficient estimation of word representation in vector space. In *Proc. of ICLR*, 2013.

[Miller and Charles, 1998] G. Miller and W. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1998.

[Mnih and Hinton, 2009] Andriy Mnih and Geoffrey E. Hinton. A scalable hierarchical distributed language model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Proc. of NIPS*, pages 1081–1088, 2009.

[Mu *et al.*, 2018] J. Mu, S. Bhat, and P. Viswanath. All-but-the-Top: Simple and Effective Postprocessing for Word Representations. In *Proc. of ICLR*, 2018.

[Muromägi *et al.*, 2017] Avo Muromägi, Kairit Sirts, and Sven Laur. Linear ensembles of word embedding models. In *Proc. of the Nordic Conference on Computational Linguistics*, pages 96–104, 2017.

[Neelakantan *et al.*, 2014] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proc. of EMNLP*, pages 1059–1069, October 2014.

[O'Neill and Bollegala, 2018] James O'Neill and Danushka Bollegala. Meta-embedding as auxiliary task regularization. In *Proc. of EACI*, 2018.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345 – 1359, October 2010.

[Passos *et al.*, 2018] Alexandre Tachard Passos, Gabriel Pereyra, Geoffrey Hinton, George Dahl, Robert Ormandi, and Rohan Anil. Large scale distributed neural network training through online distillation. In *Proc. of ICLR*, 2018.

[Pennington *et al.*, 2014] Jeffery Pennington, Richard Socher, and Christopher D. Manning. Glove: global vectors for word representation. In *Proc. of EMNLP*, pages 1532–1543, 2014.

[Peters *et al.*, 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL-HLT*, 2018.

[Poerner *et al.*, 2020] Nina Poerner, Ulli Waltinger, and Hinrich Schütze. Sentence meta-embeddings for unsupervised semantic textual similarity. In *Proc. of ACL*, pages 7027–7034, 2020.

[Polikar, 2012] Robi Polikar. Ensemble learning. In *Ensemble Machine Learning*, pages 1–34. Springer US, 2012.

[Reisinger and Mooney, 2010] Joseph Reisinger and Raymond J. Mooney. Multi-prototype vector-space models of word meaning. In *Proc. of HLT-NAACL*, pages 109–117, 2010.

[Takahashi and Bollegala, 2022] Keigo Takahashi and Danushka Bollegala. Unsupervised attention-based sentence-level meta-embeddings from contextualised language models. In *Proc. of LREC*, 2022.

[Tissier *et al.*, 2017] Julien Tissier, Christopher Gravier, and Amaury Habrard. Dict2vec : Learning word embeddings using lexical dictionaries. In *Proc. of EMNLP*, pages 254–263, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc of NeurIPS*, pages 5998–6008, 2017.

[Wang *et al.*, 2018] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proc. of BlackboxNLP*, pages 353–355, 2018.

[Winata *et al.*, 2019a] Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. Learning multilingual meta-embeddings for code-switching named entity recognition. In *Proc. of RepL4NLP*, pages 181–186, 2019.

[Winata *et al.*, 2019b] Genta Indra Winata, Zhaojiang Lin, Jamin Shin, Zihan Liu, and Pascale Fung. Hierarchical meta-embeddings for code-switching named entity recognition. *arXiv preprint arXiv:1909.08504*, 2019.

[Wu *et al.*, 2020] Xin Wu, Yi Cai, Yang Kai, Tao Wang, and Qing Li. Task-oriented domain-specific meta-embedding for text classification. In *Proc. of EMNLP*, pages 3508–3513, 2020.

[Xie *et al.*, 2019] Yuqiang Xie, Yue Hu, Luxi Xing, and Xiangpeng Wei. Dynamic task-specific factors for meta-embedding. In *International Conference on Knowledge Science, Engineering and Management*, pages 63–74. Springer, 2019.

[Yang *et al.*, 2019] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. of NeurIPS*, 2019.

[Yin and Schütze, 2016] Wenpeng Yin and Hinrich Schütze. Learning word meta-embeddings. In *Proc. of ACL*, pages 1351–1360, 2016.

[Yin and Shen, 2018] Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Proc. of NeurIPS*, pages 887–898. Curran Associates, Inc., 2018.